POLITECNICO DI TORINO

DEPARTMENT OF CONTROL AND COMPUTER ENGINEERING

Master's Degree in Computer Engineering
DATA SCIENCE

Master's Degree Thesis

Data Exploration techniques for classification analysis

Supervisors

Prof. Elena Maria BARALIS

Dott. Eliana PASTOR

Candidate

Carmen MOTTA

ACADEMIC YEAR 2020-2021

Summary

The rapid spread of Machine Learning systems in contexts where a decision is required has introduced a new challenge for designers who have to interface with new considerations inherent to fairness and the possible intrinsic bias of systems. Studies show that there may be possible unwanted biases that AI systems present against people of specific groups, often underrepresented, based on race, sex, religion, or age, among other characteristics. Also when validating a model, the overall performance may not reflect those of the smaller subsets. In this thesis new data exploration techniques will be proposed for the analysis of the classification, going to search for those subgroups in which the model is underperforming. The subdivision allows users to analyze model performance at a more granular level. We then consider the importance of moving to a use-oriented design by presenting an interactive tool to support the analysis able to guide the user in the process, allowing a greater understanding of the results obtained, offering a certain degree of interactivity with the steps carried out by increasing usability and improving the user experience also in this sector.

"Be the change you wish to see in the world"
(Mohandas Karamchand Gandhi)

To My Brother, my model
To My Mother, my strength
To My Father, my guide

Acknowledgements

This thesis marks the end of my study path, a long and demanding path, characterized by many ups and downs that often put me to a severe test. Today I can say with a smile that I have reached this important goal and I cannot help but thank all the people who have been a significant part of it, helping me, supporting me and above all strengthening me.
For this, I would like to start with a huge THANK YOU.
I thank my family for allowing me to take this path. In particular, I would like to thank my brother Michele, who every day gives me an impeccable model to follow, always making me aspire to a better version of myself. I thank my mother Stefania, who lives with me every good or bad emotion, who is able to transfer all her strength to me if I can't alone and who teaches me every day what it means to be a woman. I thank my father Antonio, my guiding spirit, the man who with his very existence allows me to believe that I can do everything and makes me who I am. I thank my grandmother with whom I spent all my childhood and who today proudly smiles at me every time she sees me.
I thank my friend Nicole Le Voci, who stole a part of my soul forever and is the extreme representation of the good you can feel for a friend, the sun of my life.
I thank my friend Nicole Farina, who always has the right word at the right time, who silently always protected me and never doubted me, making me get up every time I fell, my luck.
Also, I would like to thank my cousin Rosita, a very important person in my life, who has raised and guarded me since I was a child and even if she constantly scolds me because I never call her, she gives me a safe haven in which to take refuge.
I would also like to thank all the wonderful people I have met during this journey and who have become very meaningful to me.
I would start by thanking my boyfriend Carlos who gave me a shot of life when everything seemed flat, representing the beginning of my personal path all uphill, and giving me today the certainty of being ready for the new phase of my life.
I thank my roommate and friend, Anna, thanks to her I learned that it doesn't take years to create a strong and indissoluble bond and that you don't need to live in the same place to get excited together at any time. I would like to thank all my

colleagues who have accompanied me over the years.
I thank Giuliana and Pamela who have always had an open door for me and have overwhelmed my life like a hurricane full of happiness and smiles. I thank my IT colleagues who gave me the best group to live this experience with, Andrea, Mox, Francesco, Max and Valeria and especially Marco for having always been available and patient and for having totally understood me.
Also, I would like to thank my thesis supervisor, Professor Elena Maria Baralis who intruded me in the wonderful world of Machine Learning and for giving me the opportunity to carry out this fantastic thesis work. I would also like to thank my co-advisor, Dr. Eliana Pastor for the availability, professionalism and attention with which she followed me throughout the work.
Finally, I would like to thank myself for the commitment, the stubbornness, the seriousness and the will that I have put in every day to get to this point.
THANK YOU ALL.

Carmen Motta, Torino, April 2021

Table of Contents

List of Tables ix

List of Figures x

Acronyms xv

1 Introduction 1

2 Related Works 4
2.1 Explaining Classifications, Bias and Fairness 4
2.1.1 Slice Finder 4
2.1.2 MLCube 5
2.1.3 AI Fairness 360 5
2.1.4 FairVis 6
2.1.5 Aequitas 6
2.2 Interactive Tool to support the analysis 7
2.2.1 Interaction to understand predictions 8
2.2.2 Interaction to understand the capabilities of the classifier . . 8
2.2.3 Interaction for data analysis 9
2.2.4 Interaction to evaluate fairness and bias 9
2.2.5 Interaction to find problematic subgroups of instances 10

3 Background 11
3.1 Classification and Algorithmic Bias 11
3.2 Game Theory 14
3.3 Shapley Value 17

4 Data exploration for classification analysis 19
4.1 Metrics used in the evaluation 19
4.2 Research for representative subgroups 22
4.3 Evaluation method for frequent patterns 22

4.4 Search for a Global vision 26

5 Visual Exploration of Interactive Tool 28
5.1 Interactive Notebooks 34
5.1.1 Selection of the dataset 35
5.1.2 Analysis parameter set and Evaluation metrics selection . . 36
5.1.3 Frequent patterns extraction 37
5.2 User Interface 46
5.2.1 Front-end Environment 46
5.2.2 Back-end Environment 47
5.2.3 Graphical User Interface (GUI) 48
5.3 Description of interactive operations 54
5.3.1 Choice of the dataset 54
5.3.2 Selection of metrics 57
5.3.3 Performing the Analysis 58
5.3.4 Global Evaluation 66

6 Experiments 71
6.1 Evaluation of the results of the interactive tool 71
6.1.1 False Positive Rate (FPR) 71
6.1.2 False Negative Rate (FNR) 76
6.1.3 Accuracy 81
6.2 Comparison with FairVis 86
6.2.1 Differences on the graphical interface level 86
6.2.2 Comparison presentation of results 87
6.3 Comparison with Aequitas 94
6.3.1 Comparison of results presentation 94
6.4 Comparison with Slice Finder 99
6.4.1 Comparison of results presentation with Adult Dataset . . . 99

7 Conclusion 101
7.1 Future Work 102

Bibliography 103

List of Tables

4.1 List of metrics available in the method 20

List of Figures

3.1 Graph Counterfactual [37] 12

4.1 Frequent pattern evaluation with the overall: the second instance has a lower accuracy than the overall this indicates incorrect behavior of the model in this itemset. 23
4.2 Shapley Value graph to evaluate the local contributions of the attributes that make up the itemset. 24
4.3 Evaluation of the contributions of adding items to the itemset . . . 25
4.4 Comparison of Shapley values[44]: on the left the local contributions of the original item set attributes; on the right the local contributions of the item set with the addition of the item 26

5.1 Display Dataset selection 35
5.2 Selected Dataset's information 36
5.3 Selection of the class map for dataset uploaded by the user 36
5.4 set of analysis parameters 37
5.5 display of frequent patterns 38
5.6 display of frequent patterns with dropped column 38
5.7 Most discrepant patterns with all selected metrics 39
5.8 Most discrepant patterns with only discrepant information 39
5.9 Top K table with selection of the number of rows 40
5.10 Bar chart for Shapley Value 40
5.11 Example of Lattice Search visualization 41
5.12 Table with Corrective values 42
5.13 Comparison of the Shapley Values of two subgroups 42
5.14 Visualization of lattice search of itemset with corrective item . . . 43
5.15 Selection of values for each field with display of the selected item . 43
5.16 notice to the user item below threshold 44
5.17 Global Shapley Value 45
5.18 Global Comparison Shapley Value 45
5.19 Flask URL routing in Hello Word web application 47

5.20 First page of proposed Interactive Tool 49
5.21 About page of proposed Interactive Tool: the page is composed at the top of the navigation bar, in the center we find an automatic scrolling carousel that shows some details of the process, under a text bar containing a small summary of the problem and the proposed solution. 49
5.22 Demo page of proposed Interactive Tool: this page is an extra step proposed to the user before starting the exploration to underline some important instructions that must be taken into consideration during the analysis 50
5.23 Display examples of indications: on the left an example of a warning, to indicate to the user that he is carrying out a wrong action; on the right an example of successful action, the user is reported of the success of an operation 51
5.24 Dynamic button example display: on the left the layout of the button when it does not receive interaction; on the right the button when the mouse is positioned over it. 52
5.25 Example of page: all the pages of the tool have this scheme, navigation bar at the top for the main steps, side navigation bar for the single operations relating to the step, central content with the operation to be carried out. 52
5.26 View of Dataset Choice 54
5.27 Loading the new dataset into the application 55
5.28 Insertion of a dataset chosen by the user 55
5.29 Display of the entire dataset selected by the user 56
5.30 Setting the parameters of the analysis 57
5.31 Selection of evaluation statistics 57
5.32 Evaluation of the most discrepant patterns 58
5.33 Exploration of most discrepant patterns: pattern selection 59
5.34 Exploration of most discrepant patterns: top 20 pattern 60
5.35 Exploration of most discrepant patterns: select a range of patterns to display 60
5.36 Exploration of most discrepant patterns: Shapley Value of selected pattern 61
5.37 Exploration of most discrepant patterns: Lattice Search 62
5.38 Exploration of most discrepant patterns: Lattice Search without displaying the regulatory items 62
5.39 Search for items with a regulatory effect 63
5.40 Lattice Search for items with a regulatory effect 64
5.41 Search for information about an itemset by selecting attribute values 64
5.42 Search for information about an itemset with itemset found 65

5.43 Search for information about an itemset with itemset not found . . 65
5.44 Display of options for global evaluations 66
5.45 Evaluation of the global Shapley values for each item 67
5.46 Evaluation of the global Shapley values and group discrepancy values for each item 67
5.47 Evaluation of the global Shapley values for False Positive Rate and False Negative Rate 68
5.48 Evaluation of the most discrepant patterns with threshold for discarding itemset with irrelevant variation 69

6.1 Most discrepant patterns for Adult Dataset (FPR) 72
6.2 Selection of most discrepant pattern for Adult Dataset 72
6.3 Shapley values for most discrepant pattern for FPR (Adult Dataset) 73
6.4 Lattice Search for most discrepant pattern for FPR (Adult Dataset) 73
6.5 Regulatory item search for FPR (Adult Dataset) 74
6.6 Lattice Search of regulatory item for FPR (Adult Dataset) 74
6.7 Global Discrepancy Group for every item for FPR (Adult Dataset) 75
6.8 Most discrepant patterns for Adult Dataset (FNR) 76
6.9 Shapley values for most discrepant pattern for Adult Dataset (FNR) 77
6.10 Lattice Search for most discrepant pattern for Adult Dataset (FNR) 77
6.11 Regulatory item search for FNR (Adult Dataset) 78
6.12 Lattice Search with the effect of the regulating item for Adult Dataset (FNR) 79
6.13 Global Discrepancy Group for every item for FNR (Adult Dataset) 80
6.14 Most discrepant patterns for Adult Dataset (Accuracy) 81
6.15 Shapley values for most discrepant pattern for Adult Dataset (Accuracy) 82
6.16 Lattice Search for most discrepant pattern for Adult Dataset (Accuracy) 83
6.17 Regulatory item search for Accuracy (Adult Dataset) 84
6.18 Lattice Search with the effect of the regulating item for Adult Dataset (Accuracy) 84
6.19 Global Discrepancy Group for every item for Accuracy (Adult Dataset) 85
6.20 visualization of the tool proposed by FairVis[18] 86
6.21 FairVis[18] accuracy evaluation: we generated the groups thanks to the panel on the left and we evaluated the two sub-groups that are at the extremes compared to the average of the overall; on the left the bar graph shows in red the accuracy of the subgroup to the right of the average, in blue the accuracy of the subgroup to the left of the average. 88

6.22 Accuracy evaluation in our interactive tool: we have the table with all the subgroups automatically generated by the Frequent Pattern Mining algorithm; it is possible to sort the columns of interest in our case we evaluate the lowest accuracy values to verify how far they are from the overall. 89
6.23 Pattern Exploration: this view allows the user to evaluate, for each generated subgroup, how the group discrepancy value is produced; selecting the row of interest can display a series of details such as the lattice graph. 90
6.24 Pattern Exploration Shapley Value: selected the instance to be evaluated by pressing the next button, the user will see the bar graph that represents the Shapley values for the elements of the item set. 90
6.25 Pattern Exploration higher accuracy values. 91
6.26 Pattern Exploration Shapley Value higher accuracy values. 91
6.27 Pattern Exploration Lattice Graph:on the left the lattice graph corresponding to the instance with the highest accuracy value; on the right the lattice graph corresponding to the instance with the value of less than accuracy. 92
6.28 Evaluation of items that decrease the distance value 93
6.29 Evaluation of the False Positive Rate for sensitive characteristics (Aequitas [106]). 94
6.30 Visualizing disparities between groups in the single attribute age and race for FPR (Aequitas [106]) 95
6.31 Visualizing fairness of a single absolute group metric across all population groups (Aequitas [106]) 96
6.32 Visualizing fairness between groups in a single attribute race for all calculated disparity metrics (Aequitas [106]) 97
6.33 Representation global Shapley values for FPR and FNR 98
6.34 Evaluation of problem sections for the Slice Finder system [108] . . 99
6.35 Evaluation of problem sections with max effect size for the Slice Finder system [108] 100

Acronyms

AI
: artificial intelligence

ML
: Machine Learning

UX
: User eXperience

DOM
: Document Object Model

IML
: Interactive Machine Learning

VA
: Visual Analysis

XAI
: eXplainable Artificial Intelligence

Chapter 1

Introduction

The field of Machine Learning deals with how to build computer systems that automatically improve with experience and with researching what are the fundamental laws that govern all learning processes[1, 2, 3, 4, 5, 6]. This objective covers a wide range of learning activities, such as how to design autonomous robotic systems that learn to navigate from their own experience, how to extract information from historical medical records to learn the best treatments for future patients, and how to build search engines that adapt automatically customize to the interests of their users. Considering a particular activity T, a performance metric P and an experience type E, we can say that the machine learns whether the system reliably improve its performance P in activity T, following experience E. The different specification of T, P and E leads to a different branch of artificial intelligence such as data mining, autonomous discovery, database updating, etc [1].
Considering that Computer science is based on building machines capable of solving problems, and Statistics deals with what can be deduced from the data, with what reliability, focusing on modeling, Machine Learning can be seen as their intersection because it is based on both [1]. The difference between Computer Science and Machine Learning is that the former mainly focuses on how to program computers manually, while the latter focuses on the question of how to get computers to program themselves (from experience plus some initial structure). The difference between Statistics and Machine Learning is that the former focuses on drawing conclusions from the data, while the latter adds further questions about computational architectures and algorithms to be used to derive, store, retrieve and merge data and how to organize multiple secondary learning activities in a larger system, as well as dealing with computational tractability issues [1]. To date, the insights gained from Statistics and Computer Science are much stronger than the insights gained from Machine Learning from studies of Human Learning studies, mainly due to the weak state of our understanding of Human Learning. However, collaboration between studies of machine and human learning is growing, with

increasingly complex Machine Learning algorithms[1].
In the coming years it is reasonable to expect this collaboration between studies of Human Learning and Machine Learning to grow substantially, as they are close neighbors in the landscape of fundamental science issues.
Other fields, such as biology, economics or control theory, have an interest in finding out how systems can automatically adapt or optimize to their environment, and Machine Learning will likely to be very useful to these fields in the coming years.
One measure of advances in Machine Learning is its significant real-world applications, from Speech recognition to Computer vision to Bio-surveillance to Robot control with systems that are ever more accurate than hand-crafted programs[1].
For example , several researchers [7, 8, 9] have demonstrated the use of machine learning to acquire control strategies for stable helicopter flight and helicopter aerobatics), Accelerating empirical sciences (many data-intensive sciences now use machine learning methods to aid in the scientific discovery process).
Machine learning methods are very useful in the development of particular types of software, mainly when the application is complex for a manual algorithm design or when it requires customization based on your operating environment.
As we can see, machine learning methods play a key role in the world of computer science, within an important and growing range. Although there are software applications where machine learning may never be useful (e.g. to write matrix multiplication programs), the range where it will be used is growing rapidly in proportion to the complexity of applications, as the search for self-personalizing software increases as computers gain access to more data and as we develop increasingly effective machine learning algorithms. Shifting the focus from computer programming to how to allow them to program themselves, Machine Learning highlights the importance of designing self-monitoring systems that self-diagnose and self-repair, and approaches that train their users taking advantage of the constant stream of data provided by the program rather than simply processing it[1].
The use of artificial intelligence systems in contexts in which a decision is required has introduced a new challenge for designers, bringing new considerations to the work plan, inherent to fairness and the possible intrinsic bias of the systems. Studies have shown that there may be possible unwanted biases AI systems might have against people from specific, often underrepresented groups based on race, sex, religion or age, among other characteristics, as the ProPublica article demonstrated on the COMPAS system for the prediction of recidivism, which presents a bias towards black people in the classification [10].
The bias can be caused by several factors: it can be intrinsic in the model that encodes implicit and explicit social bias [11] (algorithmic bias), training data may not be representative either in terms of different demographic groups or within a particular demographic group, there may be an error leading to a bias in training

data labels, there may be unequal rates of labels across demographic groups, the model class may be too simple to detect relationships between characteristics for certain groups, and more [12].
The diffusion of these learning systems could be further expanded by facilitating the understanding of existing systems and introducing less experienced users into the research process. In this thesis we talk about Human-Computer Interaction (HCI), or the evaluation and implementation of computer systems interactive, that is, aimed at the use by human users.
In fact, the transition from a purely systems-oriented design to a use-oriented design becomes fundamental, so that the designer who wants to create quality products focus his activity, in a conscious and informed way, on the needs of the users of the systems he designs and on the different contexts of their use. In this thesis, new data exploration techniques for classification analysis will be presented [13]. In validating a model the overall performance may not reflect that of the smaller subsets, in Chapter 4 we will present a method [13] of evaluating the model based on the distance of performance between representative subsets of the dataset and the entire dataset, offering to the user a way to analyze model performance at a more granular level.
With a "Human-in-the-Loop" approach, we try to evaluate the behavior of the classifier in representative subgroups of the dataset, comparing a forecast with a subset or even with a single data point. In support of the work carried out an interactive tool will be presented, that can guide the user in the carried out analysis process, allowing the set of some parameters and an interactive display of the results obtained to increase usability, improving the user experience also in this sector.

Chapter 2

Related Works

2.1 Explaining Classifications, Bias and Fairness

The rapid growth of machine learning systems and their adaptability in every field of human life has led to the birth of a question that users ask themselves: "How reliable is the model?". Nowadays the understanding of a prediction of a model goes hand in hand with the search for optimization of the model itself, especially when these systems are used in more delicate areas such as medicine, finance or justice. For example Strumbelj and Kononenko [14] look for an effective general explanation method for classifiers' predictions, using only input and output of a classifier they decompose the changes in its prediction into contributions of individual feature values.This contributions correspond to known concepts from coalitional game theory. The resulting theoretical properties of the proposed method guarantee that no matter which concepts the classifier learns, the generated contributions will reveal the influence of feature values.
We cannot limit ourselves to saying that the model works, it is important to understand and be able to explain which factors directed the classifier to give a certain prediction and audit the fairness of the model to avoid implicit and explicit biases into the outputs. A necessary condition if the model used acquires a certain importance, for example to decide the risk of recidivism of a prisoner.
This chapter offers an overview of the current literature's approaches.

2.1.1 Slice Finder

Slice Finder [15] focuses on the problem of slicing data to identify subsets of validation data where the model perform poorly , this consideration is based on the fact that in model validation the overall performance of the model may not reflect that of smaller subsets and affection allows users to analyze model performance at a more granular level[15]. The point to consider is the search for problematic slices for

model validation that is to find easy-to-understand subsets of data and ensure that the model performance on the subsets is meaningful and not attributed to chance. Slice Finder discovers large possibly-overlapping slices that are both interpretable and problematic[15]. A slice is defined as a conjunction of feature-value pairs where having fewer pairs is considered more interpretable. A problematic slice is identified based on the testing of a significant difference of model performance metrics of the slice and its counterpart.
Slice Finder's approach[15] follows two paths for slice finding. The first deals with decision tree training, which has a more natural interpretation as the leaves directly correspond to slices and finds non-overlapping slices. The second deals with lattice searching, in which it is considered a large search space where the slices form a lattice, and problematic slices can overlap with one another.

2.1.2 MLCube

MLCube[16] offers users a way to define instance subsets using relational selections over features, and compute aggregate statistics and evaluation metrics over the subsets . It defines a subset as a relational selection over a feature vector table or the raw data table, and computes aggregate statistics (e.g. accuracy) for all user-defined subsets. It allows the user to view all intermediate results, allowing to define a subset not only over features, but also over data attributes, or over a combination of multiple components. MLCube[16] selects all categorical attributes and create discrete bins for selected numerical (continuous) attributes and features.

2.1.3 AI Fairness 360

AI Fairness 360 (AIF360) is concerned with the problem of fairness in machine learning models, often used to support decision making in high-stakes applications, tries to help facilitate the transition of fairness research algorithms to use in an industrial setting and to provide a common framework for fairness researchers to share and evaluate algorithms [17]. The goal of this framework is to promote a deeper understanding of fairness metrics and mitigation techniques.
The basic idea is to upload data into a dataset object, transforming it into a fairer dataset using a fair pre-processing algorithm, then learning a classifier from this transformed dataset, and obtaining predictions from this classifiers. Metrics can be computed on the original, transformed, and predicted datasets as well as between the transformed and predicted datasets.
AI Fairness[17] currently contains 9 bias mitigation algorithms divided in 3 categories, pre-processing, in-processing, post-processing, that is divided according to the position in which they can intervene in a complete machine learning pipeline.

2.1.4 FairVis

FairVis is a mixed-initiative visual analytics system that integrates a novel subgroup discovery technique for users to audit the fairness of machine learning models, allows users to explore both suggested and user-specified subgroups that incorporate a user's existing domain knowledge [18]. Users can visualize how these groups rank on various common fairness and performance metrics and the contextualize subgroup performance in terms of other groups and overall performance.
FairVis[18] generates the subgroups by executing as the first step clustering on the training dataset to find statically similar subgroups of instances. Next, it uses an entropy technique to find important features and compute fairness metrics for the clusters. Lastly, it presents users with the generated subgroups sorted by important an anomalously low fairness metrics. FairVis[18] discovers similar subgroup using similarity in the form of statistical divergence between feature distributions to find subgroups that are statistically similar.

2.1.5 Aequitas

Aequitas enables users to seamlessly test models for several bias and fairness metrics in relation to multiple population sub-groups [19]. In Aequitas bias and fairness are not absolute concepts and are not independent from the application scenario, as well as its analysis and interpretation.
Aequitas[19] provides comprehensive information on how it should be used in a public policy context, taking the resulting interventions and its implications into consideration, it also is intended to be used by policymakers, and consequently provides seamless integration in the ML workflow. Aequitas[19] can audit AI systems to look for biased actions or outcomes that are based on false or skewed assumptions about various demographic groups. Users simply upload data from the system being audited, configure bias metrics for protected attribute groups of interest as well as reference groups, and then the tool generates bias reports.

2.2 Interactive Tool to support the analysis

A machine learning system can be as efficient as it is hard to understand for the end user, especially when understanding the model is based on lines of code to examine and interpret.Considering, for example, the analysis of a model, a different representation may be relevant for the usability of the adopted algorithm. The creation of an interactive tool to support your project means shifting a part of attention to the end user who will be the user of the model if they understand the advantages, for this reason attracting the user can become a fundamental factor to keep in mind for the distribution of one's work.
In the specific case of an analysis with results, the idea is to enhance the work done by highlighting the salient features and directing the user's attention to those interesting results that have been obtained, adding explanatory graphs, tables, centering the point of analysis. Another aspect to take into consideration is the user's insertion in the analysis process, this can be done at different levels, from the choice of all the parameters used in which the user can formulate the entire analysis, to the simple request for progress if we build the tool so that the analysis is presented in steps requiring an input to pass from one step to another (for example by clicking on a button).
We can see how it is underlined in [20] that more and more researchers are realizing the importance of studying users of these systems and that interactive machine learning can facilitate the democratization of applied machine learning. Empowering end users to create machine-learning-based systems for their own needs and purposes, involving a new factor to be taken into consideration, it is necessary to understand the abilities, behaviors and needs of the end user.
Also in [21] we note how by creating correct interaction cycles we can guide the automatic learning behaviors even of users with little or no machine learning experience, through low-cost trial and error or focused experimentation with inputs and outputs. In [22] we talk about the fusion of machine learning and visual analytics as an opportunity for visual data analysis. Visual analytics leverages the cognitive and perceptual abilities of humans to enable them to explore, reason, and discover data features visually.
Machine learning leverages the computational abilities of computers to perform complex data-intensive calculations to produce results for specific questions or tasks. The discussion in Dagstuhl's seminar [22] focuses on the user's role in the process of integrating machine learning into visual analytics, identifying aspects of machine learning methods, which can be interactively controlled by the user, such as choice and parameterization of machine learning models. While some of these aspects can be automatically optimized by predefined cost functions, in many applications it is essential to allow the user to control them interactively. An example would be the view of classifiers, responding to the growing demand for

interpretable models, which lead to visualization, not only showing the data, but also an inferred classification model. This enables the use of human perceptual qualities to detect: 1) potential mis-labeling errors which might emerge as outliers, 2) noisy regions which are difficult to classify, 3) the modality of each class and 4) model overfitting effects, etc.
In the current literature we find many examples of interactive tools to support machine learning systems.

2.2.1 Interaction to understand predictions

Prospector [23] is an interactive visual analytics system that offers a way to understand how features affect the prediction overall. Prospector[23] helps analysts better understand predictive models interactively by leveraging the concept of partial dependence, a diagnostic technique used to determine how features affect the prediction, and makes this technique fully interactive, supports localized inspection, so users can understand why certain data results in a specific prediction, and even lets users hypothesize new data by changing values and observing how the predictive model responds.
Rivelo [24] is a visual analytics interface that enables analysts to understand the causes behind predictions of binary classifiers by interactively exploring a set of instance-level explanations, these explanations are model-agnostic, treating a model as a black box, and they help analysts in interactively probing the high-dimensional binary data space for detecting features relevant to predictions.

2.2.2 Interaction to understand the capabilities of the classifier

[25] presents an interactive tool based on learning the kernel and hyperparameters for multiclass classification that leverages human guidance, the method enables people to prune the model space via interactive exploration, reducing computational needs. Starting with an initial model, users can interact with a visual representation of a leave-one-out confusion matrix, allowing them to search among a space of models to identify a model whose cross-validation performance is favorably aligned with the desired output.
The key idea is to harness user interactions to explore the space of solutions without cross validating the entire space in an exhaustive manner. By visualizing the possible solutions and guiding the search, users can both gain a sense of the capabilities of the classifier and choose a model aligned with his goal.
Squares [26] is an interactive performance visualization for multiclass classification problems. Squares displays information used to derive several common performance

metrics and helps practitioners prioritize efforts in debugging performance problems while supporting direct access to instances. It allows users to click on boxes, strips, or stacks to reveal corresponding instances or groups of instances in an adjacent table. The instances can be color coded by their true label and performance issues are indicated by the arrangement of colored points in the display (e.g., a mix of colored points may indicate poor separability of certain classes). Users can also click on individual points to view the corresponding data instances.

2.2.3 Interaction for data analysis

InsightsFeed [27] was developed for analyzing large Twitter datasets using the PVA paradigm. Progressive visual analytics (PVA) is a data analysis system that deliver improving estimates of the results of computations have been introduced, these systems involve the analyst during long computational processes allowing interactive exploration, for example by filtering data or changing the parameters of the computation and progressively visualizing their intermediate results. Advantages of PVA include (1) reduced latency, (2) better transparency of how the computational methods work, and (3) support for early decision-making, either for making a final decision or for terminating misguided analyses early. InsightsFeed[27] uses the traditional visual analytics pipeline and incorporate intermediate results and feedback about progress from long computations, while providing easy controls to change the parameters of the progressive computations as well as supporting interactive filtering of visualized data. The InsightsFeed[27] interface helps the data analyst understand the tweets, sentiments, as well as keywords discussed within tweets that are visualized using multidimensional projection algorithms to create a semantic map.
CueFlik [28] is a system developed to support Web image search,in which it is shown that well designed interactions can significantly impact the effectiveness of the interactive machine learning process. CueFlik[28] allows end-users to interactively define visual concepts (e.g., “product photos”, “pictures with quiet scenery”, “pictures with bright psychedelic colors”) for re-ranking web image search results. End-users train CueFlik[28] by providing examples of images with and without the desired characteristics. These examples are used to learn a distance metric as a weighted sum of component distance metrics (including histograms of pixel hue, saturation, luminosity, edges, global shape and texture).

2.2.4 Interaction to evaluate fairness and bias

FairVis [18] offers an interactive visual interface to help users explore the fairness of their machine learning models and discover potential biases.

AI Fairness 360 [17] is an open source toolkit that brings value to diverse users and practitioners (2.1.3). For fairness researchers, it provides a platform that enables them to: 1) experiment with and compare various existing bias detection and mitigation algorithms in a common framework, and gain insights into their practical usage; 2) contribute and benchmark new algorithms; 3) contribute new datasets and analyze them for bias. For developers, it provides: 1) education on the important issues in bias checking and mitigation, 2) guidance on which metrics and mitigation algorithms to use; 3) tutorials and sample notebooks that demonstrate bias mitigation in different industry settings; and 4) a Python package for detecting and mitigating bias in their workflows.

2.2.5 Interaction to find problematic subgroups of instances

MLCube Explorer [16], an interactive visualization tool for exploring machine learning results using MLCube (see 2.1.2). MLCube Explorer[16] allows users to visually explore aggregate statistics over subsets of data instances and interactively drill down into models. This enables users to find interesting patterns between features and model results, leading to discovering insights that help them understand the mechanisms of the models and further improve their performance.
Slice Finder [15] is an interactive framework for identifying such slices using statistical techniques (2.1.1), the interface consists of: a scatter plot that shows the (size, effect size) coordinates of all slices. This gives an overview of the top-k problematic slices, which allows the user to quickly browse through large and also problematic slices and compare slices to each other. The user can view the slice description, size, effect size, and metric (e.g., log loss), can select a set of slices and view their details on a table; on the table view, the user can sort slices by any metrics on the table.

Chapter 3

Background

This chapter presents a background of basic concepts that will be used in the description of the analysis carried out in this thesis.

3.1 Classification and Algorithmic Bias

Classification is a technique used in supervised learning where the goal, based on the analysis of previously labeled data, is to be able to predict the labeling of future data classes. Labels are unordered discrete values that can be considered to belong to a group of a class. The algorithm is instructed by the supervisor to recognize the categories through a series of practical examples (dataset training). In each example the machine is supplied with: the descriptive variables of the environment (x) a label to indicate the desired result (y) that is, the class to which the example belongs. The system processes the examples in search of a general classification rule called a model. Once the model has been built, the machine uses it to classify the new instances, based on the observations made on the training set [29, 30, 31]. A classification model generated via a learning algorithm must be able to adapt correctly to the input data, but also and above all be able to correctly predict record class labels that never has seen before. That is, the key objective of the learning algorithm is to build models with good generalization skills. Given the growing popularity of machine learning systems in many different areas of practice in our society, the evaluation of the fairness of the model is becoming increasingly important. This is because despite the benefits that algorithmic systems can make, models can reflect, inject or exacerbate implicit and explicit social prejudices in their outputs, disregarding some demographic subgroups [32, 33].
Find out which ones bias introduced a machine learning model is one great challenge, thanks to the numerous definitions of fairness and the great number of potentially imparted subgroups [17, 18, 19, 34, 35, 36].

Narayanan has described at least 21 mathematical definitions of equity from the literature [21], these are not just theoretical differences in how to measure fairness; different definitions produce entirely different outcomes, we can divide all the definitions of equity into 5 large partitions[37]: 1) Group Fairness, 2) Individual fairness,3) Counterfactual fairness, 4) Preference-based fairness, 5) Fairness through unawareness. We present the definitions of Group fairness and Counterfactual fairness.

Definition 3.1.1 (Group Fairness) *A classifier C satisfies the definition of group fairness if subjects in both protected and unprotected groups have equal probability of being assigned to the positive predicted class[38].*

$$P(R = +|A = a) = P(R = +|A = b).\forall a, b \in A$$

Definition 3.1.2 (Counterfactual Fairness) *A classifier C satisfies the definition of counterfactual fairness if*

$$P(C_{A\leftarrow a}(U^2) = y|X = x, A = a) = P(C_{A\leftarrow a'}(U) = y|X = x, A = a).$$

That is, given a set of attributes (e.g. level of education, type of crime, drug problems and protected attribute A = ethnicity) and a result $\hat{Y}$ to be predicted (e.g. relapse), a graph is counterfactually correct if the ethnicity is not directly connected to $\hat{Y}$ through others attributes. Intuitively, this means that a decision is right comparisons of an individual if it is the same in the (i) real world and in the (ii) world counterfactual in which the individual belonged to a different demographic group (i.e white instead of black) [34].

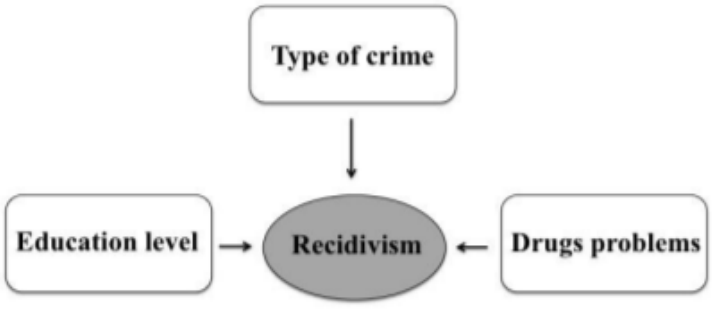


Figure 3.1: Graph Counterfactual [37]

A great difficulty in machine learning fairness is the mathematical formulation, it is impossible to satisfy all the definitions of equity simultaneously when the populations have different base rates.This incompatibility between fairness metrics was formalized by the impossibility theorem for fair machine learning [18]. Two papers [39, 40] simultaneously proved that if groups have different base rates in their labels, it is statistically impossible to ensure fairness across three base fairness metrics — balance for the positive class, balance for the negative class, and calibration of the model. Data scientists must therefore decide which fairness metrics to prioritize in a model and how to make trade-offs between metric performance.
An example of these considerations can be seen in the recidivism prediction tool COMPAS , a system that is used to predict the risk of letting someone go on bail. The ProPublica article [10] showed that in assigning the criminals' risk scores to determine their likelihood of recidivism, COMPAS was biased to give higher risk and therefore predict a higher rate of recidivism for black defendants compared to other races. The probability of a non-recidivating black defendant being assessed as high-risk is nearly double that of white defendants. Similarly, the likelihood of a recidivating black defendant being assessed as low risk is nearly half that of white respondents[10]. In technical terms, indicate that the COMPAS tool has significantly higher false positive rates and lower false negative rates for black defendants than for white defendants [36].
There have been various solutions proposed for addressing algorithmic bias in machine learning across the entire model training pipeline. These range from techniques for obfuscating sensitive variables in training data [41], to new regularization parameters for training [42] and post-processing outcomes by adding noise to predictions [43]. While these can help balance certain inequities, the impossibility theorem dictates that hard decisions will still have to be made about which fairness metrics are the most important for each problem.

3.2 Game Theory

In this section will be tell some history of game theory, Before moving on to the concept of Shapley Value [44](even when we are not speaking of the Shapley value', the history of game theory is inextricably connected with other aspects of Shapley's work). Although game-theoretic ideas can be traced earlier, much of the modern theory of games traces its origins to the 1944 book by John von Neumann and Oskar Morgenstern, *Theory of games and economic behavior* [45].
In seeking a way to analyze potentially very complex patterns of strategic behavior, their approach was to "divide the difficulties", by finding simple models of the strategic environment itself. Their first step was to find a way to summarize each alternative facing an individual decision maker by a single number. Their contribution was to specify conditions on an individual's preferences over possibly risky alternatives sufficient so that his choice behavior could be modeled as if, faced with a choice over any set of alternatives, he chose the one that maximized the expected value of some real-valued function, called his utility function.
In this way, a complex probability distribution over a diverse set of alternatives could be summarized by a single number, equal to the expected utility of the lottery in question. Once the alternatives are reduced facing each individual to a numerical description, von Neumann and Morgenstern proceeded to consider (among other things) a class of games in which the opportunities available to each coalition of players could also be described by a single number[45].
They considered cooperative games in characteristic function form (now sometimes also called "coalitional form") defined by a finite set $\mathcal{N} = \{1, \cdots, n\}$ of players, and a real-valued "characteristic function" v, defined on all subsets of $\mathcal{N}$ ($with\ v(\varphi) = 0$). The interpretation of v is that for any subset S of N the number $v(S)$ is the worth of the coalition, in terms of how much "utility" the members of S can divide among themselves in any way that sums to no more than $v(S)$ if they all agree. The only restriction on v that von Neumann and Morgenstern proposed was that it be superadditive; that is, if S and T are two disjoint subsets of N, then $v(S \cup T) \geq v(S) + v(T)$.
This means that the worth of the coalition $S \cup T$ is equal to at least the worth of its parts acting separately. The characteristic function model assumes the following things about the game being modeled. First, utility can be embodied in some medium of exchange "utility money" that is fully transferable among players, and such that an additional unit of transferable utility always adds a unit to any player's utility function. Second, the possibilities available to a coalition of players can be assessed without reference to the players not included in the coalition. Third, a coalition can costlessly make binding agreements to distribute its worth in any way agreed to by all the members, so it is not necessary to model explicitly the actions that players must take to carry out these agreements. In recognition of

the importance of the assumption that utility is transferable, these games are sometimes called transferable utility (TU) games. Although these simplifying assumptions are obviously substantial, the characteristic function model has proved to be surprisingly useful as a simple model of strategic interaction.
At the foundation of the theory of games is the assumption that the players of a game can evaluate, in their utility scales, every "prospect" that might arise as a result of a play. In attempting to apply the theory to any field, one would normally expect to be permitted to include, in the class of "prospects," the prospect of having to play a game. The possibility of evaluating games is therefore of critical importance.
Considering three essential assumptions: (a) that utility is objective and transferable; (b) that games are cooperative affairs; (c) that games, granting (a) and (b), are adequately represented by their characteristic functions, we describe below the basic concepts and definition of coalition game theory[44].

Definition 3.2.1 (Coalition Game [44]) *A coalitional form game is a tuple* $\langle N, v\rangle$*, where* $\mathcal{N} = \{1,2,\cdots,n\}$ *is a finite set of n players, and* $\mathrm{v} : 2^N \to \Re$ *is a characteristic function such that* $\mathrm{v}(\varphi) = 0$*. Subsets of N are coalitions and N is referred to as the grand coalition of all players. Function* v *describes the worth of each coalition. We usually assume that the grand coalition forms and the goal is to split its worth* $\mathrm{v}(N)$ *among the players in a "fair" way. Therefore, the value (that is, solution) is an operator* φ *which assigns to* $\langle N, v\rangle$ *a vector of payoffs* $\varphi(\mathrm{v}) = (\varphi_1, \cdots, \varphi_n) \in \Re^n$*. For each game with at least one player there are infinitely many solutions, some of which are more "fair" than others. The following four statements are attempts at axiomatizing the notion of "fairness" of a solution* φ *and are key for the axiomatic characterization of the Shapley value.*

Axiom 1 *The first axiom ("symmetry")defines that value is a property of the abstract game.*
If for two players i *and* j $\mathrm{v}(S \cup \{\mathrm{i}\}) = \mathrm{v}(S \cup \{\mathrm{j}\})$ *holds for every S, where* $S \subset N$ *and* i, j $\notin S$ *,then* $\varphi_i(\mathrm{v}) = \varphi_j(\mathrm{v})$*.*

Axiom 2 *The second axiom ("efficiency") defines that the value represents a distribution of the full yield of the game. This excludes, for example, the evaluation* $\varphi_i[\mathrm{v}] = \mathrm{v}((i))$ *in which each player assumes that the others will all cooperate against him.*

$$\sum_{\mathrm{i}\in N} \varphi(\mathrm{v}) = \mathrm{v}(N)$$

Axiom 3 *The third axiom ("dummy") defines that a player's value is zero if the value of a coalition never changes when he joins it.*
If $\mathrm{v}(S \cup \{\mathrm{i}\}) = \mathrm{v}(S)$ *holds for every S, where* $S \subset N$ *and* $\mathrm{i} \notin S$*, then* $\varphi_i(\mathrm{v}) = 0$*.*

Axiom 4 *The fourth axiom ("law of aggregation") defines that if two independent games are combined, then their values must be added player by player.*
For any pair of games v, w*:*

$$\varphi(\mathrm{v}+\mathrm{w}) = \varphi(\mathrm{v}) + \varphi(\mathrm{w}), \ \textit{where} \ (\mathrm{v}+\mathrm{w})(sS) = \mathrm{v}(S) + \mathrm{w}(S) \ \ \textit{for all } S$$

Theorem 3.2.1 *For the game* $\langle N, v\rangle$*, there exists a unique solution* φ*, which satisfies axioms 1 to 4 and it is the Shapley value:*

$$Sh_i(\mathrm{v}) = \sum_{S \cup N \setminus \{i\}, s=|S|} \frac{(\mathrm{n}-s-1)!s!}{\mathrm{n}!} \left(\mathrm{v}\left(S \cup \{\mathrm{i}\} - \mathrm{v}(S)\right)\right), \qquad \mathrm{i} = 1, \cdots, n.$$

3.3 Shapley Value

Mentioning the article [46] the Shapley Value finds a solution to the question of how to obtain a personal attribution (of payoff) to each players starting from the value of a subsets of the player set.
Shapley[44] proposed an answer to this question, which is based on the idea of defining a "value" for each player in the game, in order to evaluate whether it is worthwhile to participate. Let's consider a game with a set $N = \{1, \cdots, n\}$ of players, the value is a vector of n numbers representing the value of the game in each of its n positions. The result of Shapley[44] is that the axioms defined in the previous section uniquely determine this payout vector for each game. Using the terminology of game theory, let a coalitional game be defined by a pair (N, v), where $N = \{1, \cdots, n\}$ is the set of all players and $v(S)$, for every $S \subseteq N$, is a real number associating a worth with the coalition S, such that $v(\varphi) = 0$, this type of game is most commonly referred to as a coalitional game with transferable payoff. A payoff profile of a coalitional game is the assignment of a payoff to each of the players. A value is a function that assigns a unique payoff profile to a coalitional game. It is efficient if the sum of the components of the payoff profile assigned is $v(N)$.That is, an efficient value divides the overall game's worth between the different players .
A value that can determine the importance of the different actors can be used as a basic concept to be able to quantify the contributions of the individual elements of the system. In game theory the definite value for this type of coalitional game is the Shapley value (Shapley, 1953), defined as follows.
Let the marginal importance of player i to a coalition S, with $i \notin S$, be

$$\Delta_i(S) = v(S \cup \{i\}) - v(S)$$

Then, the Shapley value is defined by the payoff

$$\gamma_i(N, v) = \frac{1}{n!} \sum_{R \in \Re} \Delta_i(S_i(R))$$

of each player i ∈ N, where R is the set of all $n!$ orderings of N and $S_i(R)$ is the set of players preceding i in the ordering R. To understand the significance of the Shapley value we consider all the players arranged in a certain order, with all orders equally probable. Then $\gamma_i(N, v)$ is the expected marginal importance of player i to the set of players who precede him. The Shapley value is efficient since the sum of the marginal importance of all players is $v(N)$ in any ordering.
The Shapley value[44] has a wide range of applicability as illustrated in the survey paper by Moretti and Patrone (2008)[46], which is dedicated entirely to this unique

solution concept and its applications. An example of the use of Shapley value in machine learning can be found in the work of Keinan et al. [47].
We find another example of the use of Shapley Value in the context of *eXplainable Artificial Intelligence (XAI)* applied to Machine Learning Classification problems. XAI is a branch of artificial intelligence that deals with interpretability, that is, it tries to give a reasonable explanation to the results of a model [48]. The example in question is *Shapley Additive exPlanations(SHAP)* , which is a method that uses the optimal Shapley values deriving from game theory [49] with a kernel-based estimation approach. SHAP tries to explain a single forecast by calculating the contribution of each feature contributes to the resulting forecast. One of the SHAP approaches is the SamplingExplainer which calculates the SHAP values under the assumption of functionality independence by extending the algorithm proposed by Strumbelj and Kononenko [14]. If interested in learning more about the topic of shapley value, refer to [44, 50, 46].

Chapter 4

Data exploration for classification analysis

The approach presented [13] is based on the idea of searching for representative subgroups of a chosen dataset in which the model performance is inefficient and incorrect, to do this we describe below the techniques used to find the subgroups and how we have defined and evaluated the discrepancy between the ideal and the real result obtained. In this section we will deal with the theoretical basis on which the interactive tool has been built, highlighting the innovative aspects of the proposed solution which will be made even clearer in the graphic representation.

4.1 Metrics used in the evaluation

During the description of the method [13] we will see, in addition to the operation, the variety of parameters that the user can modify interactively, formulating new hypotheses and being able to explore the results obtained at a high level. Among the various choices proposed to the user there is also the choice of the metrics to be used in the validation of the model and the list of possible values is considerable.

False Positive Rate (FPR)
False Negative Rate (FPR)
Accuracy
Positive Rate
Negative Rate
|FPR|
|FNR|
|Accuracy|
Accuracy Subgroup Fairness(ACsf)
Statistical Parity Subgroup Fairness (SPsf)
False Positive Subgroup Fairness (FPsf)
False Negative Subgroup Fairness (FNsf)

Table 4.1: List of metrics available in the method

The metrics, in the context of classification, are used to quantify the goodness of the model in a more or less detailed way, depending on the metric[29, 30].
A common metric used in these cases is *Accuracy*, which measures how often the algorithm correctly classifies a given point; defined as the number of correct predictions on all predictions, too low an accuracy makes the model unsuitable or not at all; moreover, accuracy does not distinguish between false positives and false negatives and for this reason we provide other metrics to quantify them. Starting from the relationships indicated in the confusion matrix we can define the following metrics[51, 52, 53].
Another choice is the *False Positive Rate (FPR)* metric, a precision metric that can be measured on a subset and indicates the percentage of negative cases mistakenly identified as positive cases, it is a fraction of negative cases mistakenly identified as positive among all. the negative cases.
The metric *False Negative Rate (FNR)*indicates the false negative rate, it is a percentage of positive cases mistakenly identified as negative out of all positive cases. Of the metrics listed above, we also propose the version in which we consider the absolute value.
The metric *Positive Rate* indicates the rate of positive cases, it is the percentage of potential cases correctly identified as positive.
The metric *Negative Rate* indicates the rate of negative cases, it is the percentage of negative cases correctly identified as negative.
In evaluating the result of a classifier it is important to consider the hypothesis that the forecast may be influenced by some discrete random variables that encode

sensitive characteristics present in the dataset (such as example ethnicity or gender). If the prediction is independent of these sensitive characteristic, the algorithm is defined fair. The fairness of classification systems is a very recent issue that is still under development. Despite the multitude of definitions of fairness in the literature [38], there are three criteria of fairness that serve to verify this situation.
Independence: the sensitive characteristics are statistically independent of the prediction, the prediction of an individual, with different sensitive characteristics, in one group or another, is equally probable[51].
Separation: the sensitive characteristics are statistically independent of the forecast given the value of the target, i.e. the probability of being identified in a certain group is the same for two individuals with different sensitive characteristics because they actually belong to the same group, given the value of the same target[51].
Sufficiency: the sensitive characteristics are statistically independent from the value of the target given by the forecast, indicating that the probability of two individuals, with different sensitive characteristics, of being effectively identified in each of the groups is equal since they were expected to be of the same group[51].
Starting from these definitions we can introduce other metrics to evaluate the fairness of the algorithm, which we can divide into two groups: the metrics based on the expected result and the metrics based on the expected and actual results.
Among the metrics based on the expected result, we considered Statistical Parity[51], a definition that is satisfied if the cases considered of protected and unprotected groups have the same probability of being identified as positive.
Finally, we have the metrics based on the expected and actual results and among these, we considered the *Predictive Equality* (also defined as *False positive error rate balance*), a definition satisfied if the cases considered of the protected and unprotected groups have the same FPR[51].
The *Equal Opportunities* (also referred to as *False negative balance of the error rate*), and defining cases considered fulfilled if the protected groups and unprotected have FNR equals[51].
The *Overall accuracy equality* is satisfied if the case considered in the protected groups and unprotected has the same accuracy of prediction[51].

4.2 Research for representative subgroups

The first step of the analysis focuses on searching for those representative subgroups of the dataset in which the model demonstrates incorrect behavior. There are two approaches to search for subgroups, the first is aimed at the user by requesting to indicate the attributes of interest or their values, the second instead is based on automatic detection of subgroups and this is the approach followed in this analysis. In our case we make a premise, that is, the choice of subgroups is limited to those elements that have a certain frequency in the data set that have a greater number of instances and therefore a greater influence, unlike infrequent subgroups that are less relevant and whose measurements could be affected by statistical fluctuations; we, therefore, focus on the frequent subgroups.
Frequent patterns are defined as those datasets that appear in a dataset with a frequency not less than a threshold (indicated in our interactive tool as *"threshold support"*) specified by the user.
In the world of Data Mining, this process is carried out by algorithms defined *Frequent Pattern Mining* which has the objective of extracting sets of frequent elements from a database; in our system, it is possible to use any Frequent Pattern Mining algorithm considering that the performance of our model will depend on the efficiency of the chosen algorithm.
Frequent Pattern Mining algorithms require the presence of discrete data, therefore continuous attributes (if any) must be discretized. The datasets proposed in the tool are already discretized, any new datasets chosen by the user must be inserted already discretized.

4.3 Evaluation method for frequent patterns

Once we have determined the frequent subgroups we explore in this section what considerations we made to evaluate the pattern in these sets.
The idea starts from considering the fact that an underperforming subgroup must present a significant difference with respect to the behavior of the whole set, and starting from this thesis we have introduced as a comparison value the difference between the statistics computed on the item and the statistics computed on the entire dataset, which for simplicity we will refer to in the course of the explanation as a *group discrepancy* [13].

Definition 4.3.1 (Group Discrepancy) *We consider a set of arbitrary itemset* I *in the dataset* D *and denote by* p *the value of the performance computed on the element. We define the group discrepancy as:*

$$\Delta_p = p\,(I) - p\,(D) \tag{4.1}$$

The value of the group discrepancy is computed in the process of extracting the frequent subgroups to improve the estimate, once it is determined that the item considered at that moment belongs to the frequent patterns, i.e. its support value is above the indicated threshold from the user, the system returns the itemset indicating the composition of the elements and its group discrepancy value. We thus obtain a table of values in which it is possible to determine which itemsets have a greater group discrepancy value, which can also be interpreted as which groups deviate most from the correct behavior.
To give an example, let's consider accuracy as a statistic and take COMPAS as a dataset. COMPAS [54] is a decision support system used by US courts to assess the probability of a defendant becoming recidivism. The tool assigns defendants scores indicating the probability of recidivism based on more than 100 factors, including age, gender and criminal history. These scores are used to decide whether to release the accused on bail or detain him pending trial. As we can see in the Figure 4.1, the first line indicates the overall and it is the instance that we will use to make the comparison with all the others (accuracy = 0.63); in the second line, however, the item set consisting of age_cat = Less than 25, race = African-American, length_of_stay = <week has a lower accuracy (accuracy 0.53) with a group discrepancy of -0.098, this indicates that considering that subgroup the model is less accurate than the entire data set.

	support	itemsets	tn	fp	fn	tp	length	support_count	d_fpr	d_fnr	accuracy	d_accuracy
0	1.000000	()	3066	297	1962	847	0	6172.0	0.000000	0.000000	0.633992	0.000000
151	0.101264	(age_cat=Less than 25, race=African-American, length_of_stay=<week)	218	48	242	117	3	625.0	0.092137	-0.024374	0.536000	-0.097992
104	0.134964	(sex=Male, age_cat=Less than 25, length_of_stay=<week)	304	57	327	145	3	833.0	0.069581	-0.005673	0.539016	-0.094977
141	0.107583	(sex=Male, age_cat=Less than 25, race=African-American)	185	52	251	176	3	664.0	0.131095	-0.110647	0.543675	-0.090318

Figure 4.1: Frequent pattern evaluation with the overall: the second instance has a lower accuracy than the overall this indicates incorrect behavior of the model in this itemset.

It is a very useful indication in the debugging of the models, with the calculation of an estimate of the group discrepancy for the frequent subgroups we are able to obtain an indication of the behavior of the classification model. In our analysis is treated as a black box without having access to the operation internal (defined agnostic approach), proving to be an effective method for understanding the model or even for verifying a significant deviation from the general behavior for sensitive attributes.
Once the frequent itemsets have been obtained, each characterized by its group discrepancy value, the next step tries to explore in more detail for each subgroup what individual contributions the individual elements of the particular itemset have had to determine the corresponding group discrepancy value, to make this we use the method described in chapter 3 of the Shapley Value[44].
Thanks to the Shapley Value[44] we have the way to individually determine which attribute values have influenced the most in achieving a high group discrepancy value, at a higher level we can say that the presence of that attribute value during the classification process makes so that the prediction has a greater probability of being incorrect. Taking up the example considered in Figure 4.1 we can evaluate, thanks to the Shapley Value[44], the local contributions that the attributes that compose it have.

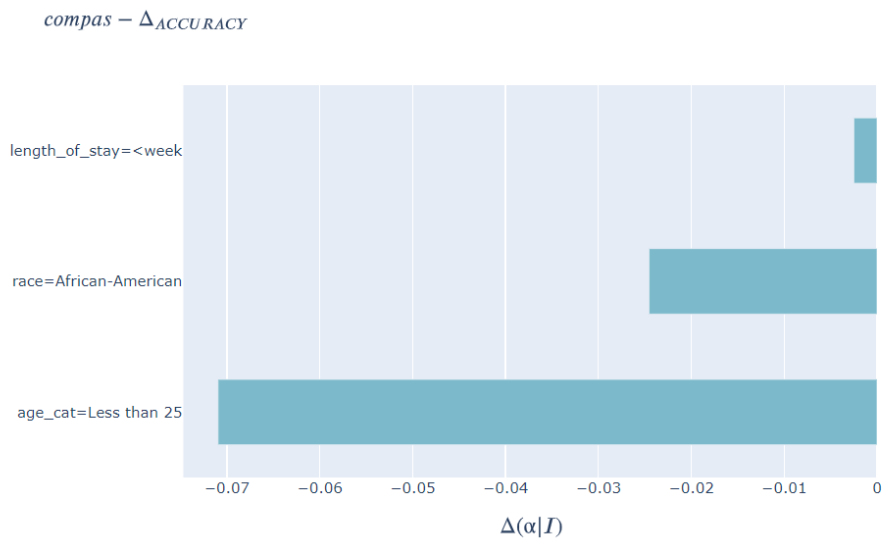


Figure 4.2: Shapley Value graph to evaluate the local contributions of the attributes that make up the itemset.

Noting that the most contributing attribute is age _cat = Less than 25 (followed by race = African-American), i.e. the presence of this attribute tends to decrease accuracy.
Summarizing the steps performed up to this point, we extracted the frequent patterns, we calculated the group discrepancy value, and thanks to the Shapley values[44] we quantified the single contribution of the attribute values that make up the item set, identifying which of them collaborate in a greater discrepancy value, noting however that if an itemset is a subgroup of another itemset this does not guarantee a smaller group discrepancy value, and from this consideration, we introduce the next step concerning the search for those items that contribute to lower the value group discrepancy when added to an itemset.
We always take accuracy as an example and consider the first instance represented in Figure 4.2, we can see how adding the item "priors _count = 0" to the item set consisting of sex = Male, race = African-American leads to an increase in accuracy from -0.038 to 0.015 by adjusting the value by an amount equal to 0.023.

	item i	S	v_S	v_S+i	t_value_corr	corr_factor
290	(priors_count=0)	(sex=Male, race=African-American)	-0.038790	0.015379	2.540020	0.023412
85	(race=African-American)	(c_charge_degree=M)	0.030858	-0.013972	2.432637	0.016886
296	(priors_count=[1,3])	(c_charge_degree=M, length_of_stay=<week)	0.044867	-0.007410	2.381362	0.037457
244	(race=African-American)	(c_charge_degree=M, length_of_stay=<week)	0.044867	-0.009307	2.676936	0.035560
291	(race=African-American)	(sex=Male, priors_count=0)	0.076534	0.015379	2.776289	0.061155
97	(race=African-American)	(priors_count=0)	0.092626	0.039767	2.818701	0.052860
259	(race=African-American)	(length_of_stay=<week, priors_count=0)	0.101443	0.043215	2.955671	0.058229

Figure 4.3: Evaluation of the contributions of adding items to the itemset

We explored all the frequent itemsets by cyclically adding a new attribute taking into account all the possible combinations with the single values that the attribute can take. We rephrased the calculations to determine which attributes exhibited this regulatory behavior of the group discrepancy value, evaluating not only the difference concerning the item-exempt from the attribute but also managing to quantify the adjustment that the added attribute provided. Finally, we tapped again on the power of the Shapley values[44] to obtain a comparison between the contributions of the initial item set and those governed by the added attribute.

Let's consider the instance examined in Figure 4.3, the following image shows the comparison between the two Shapley Values and we can see the positive local contribution that the item priors _count = 0 adds with respect to the other negatives, helping to increase the accuracy of the itemset.

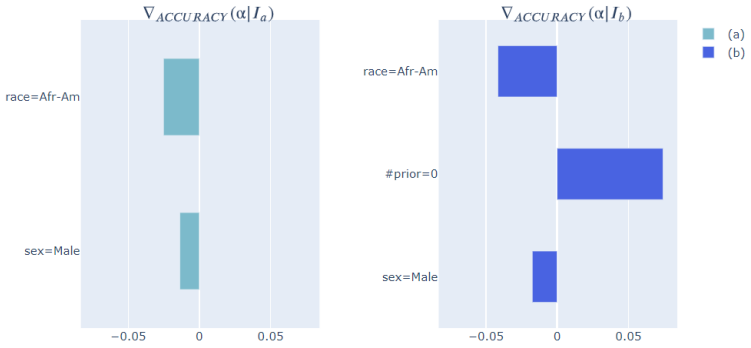


Figure 4.4: Comparison of Shapley values[44]: on the left the local contributions of the original item set attributes; on the right the local contributions of the item set with the addition of the item

4.4 Search for a Global vision

The exploration of the group discrepancy of frequent itemsets has brought to light the consideration that subgroups of an itemset do not involve lower group discrepancy values, hence the idea of a global exploration of the value of group discrepancy.
All the computations carried out up to this point evaluate the individual group discrepancy which, as we have seen, depends on how many instances have a certain configuration and consequently on how the data were collected.
This means that if in the dataset there are multiple instances of a given set rather than another for how the data set is sampled we get an influence on the estimation of the group discrepancy of the single itemsets. The evaluation is limited to the specific case examined and does not reveal the correlations between several elements, we are talking about a type of comparison between the single and the total.
We wondered if it was possible to extend the concept of group discrepancy towards a result that would give a more global vision.
In detail we tried to evaluate the contribution that a single attribute value had on

the group discrepancy if added to other itemsets thus obtaining an overall estimate of the effect of that attribute. We obtain this result using Shapley's collective game theory based on game theory, in this way we can estimate the contribution of the group discrepancy of a correlated element with all other elements, and with the overall result, we can determine if the discrepancy value of group increases as the attribute is added to the other itemset.
The global discrepancy result is a better estimate of the effect of an attribute on the discrepancy and is more robust when variations occur in the data story, and allows us to recognize which attributes tend to bend the classification towards one class or another.

Chapter 5

Visual Exploration of Interactive Tool

In this Chapter we deal with another important aspect to consider when conducting an analysis, the comprehension. The analysis represents an investigation, a detailed study of a procedure (or system), carried out through the examination of its individual elements and their interrelationships, in order to make it more rational and efficient, or to discover its malfunctions. The ultimate goal of an analysis concerns the presentation of the results obtained and the explanation of how the analysis was carried out. The analyst, often, does not pay the right attention to this last phase, focusing mainly on the formulation, algorithms and numerical results, as the recipients of the analysis are usually scholars in the sector who can understand the results obtained, also through a laborious explanation of the work or with the simple presentation of the numerical results without a pre-exposure of the basic concepts. All this implies a limitation of the diffusion of the work, limited to the field of study and only to expert users able to understand, evaluate and manipulate the algorithms involved in the analysis. Instead, a well thought out representation could facilitate the understanding process leading to a greater diffusion of the results obtained, in some cases a good graphic representation of schemes and results is sufficient (for example the use of tables, bar graphs, etc.) when the analysis does not present significant parameters for the result, in others, if the analysis allows it, the representation can be brought towards a greater direct involvement of the user in the analysis process itself, making him almost the author of results [55, 56].
By allowing the user to interact with the system, we raise the level of interest in the system and lead the user to train together with the system, obtaining greater understanding, usability and perception of what is being done and what has been achieved to expert users and in this way we are able to introduce even less

experienced users into the recipients of the work. This is achieved by building an interactive system that guides the user from start to finish, in the analysis process and in the choice of parameters manipulating the result. An efficient analysis is the one that leads to interesting results, by playing on the perception of the recipient we are able to involve them emotionally in the work, to make them participate and therefore to obtain greater understanding, interest and consequently greater distribution.
The diffusion of the use of an algorithm is strongly influenced by the perception that people have of it, regardless of its performance, this is because we are not able to establish with certainty how people perceive the decisions made by algorithms with respect to the decisions made by humans. The strength of algorithms lies in enabling efficient, optimized, and data-driven decision-making, and this aspect leads to increasing adoption of algorithms for managerial and organizational decisions. Although algorithms are increasingly efficient there is still a tendency to doubt the decision-making results that do not come from people, and this influence the perception of the decisions that are made[57]. Misperceptions of algorithmic decisions can in turn influence people's trust and attitudes toward the use of algorithms by decreasing their distribution. People form different personal theories about how algorithms work, regardless of how algorithms actually work [58].
Interactive Machine Learning (IML) is a branch of Machine Learning that combines human perception and intelligence with the computational power and speed of computers. The interactive process is designed to involve input from the user, without requiring the in-depth knowledge that might be necessary to work with more traditional machine learning techniques.
Under the IML process, non-experts can use their knowledge of the domain and datasets to find patterns of interest or develop complex data driven applications. The IML process involves the user involvement in the training process using for example human input in the example selection, creation and labelling process. [59] and [60] demonstrated that typical machine learning tasks could be designed by including human input, and over the past decade and a half the IML process has seen increasing attention within the HCI community.
The user providing input to the IML system need not possess any deep understanding of the models with which they are interacting. *Interactive Machine Learning* is based on the construction of interaction systems in which a user or user group iteratively operates a mathematical model to describe a concept through iterative cycles of input and review. Model refinement is driven by user input which can be designed in many forms, such as providing indicative samples, describing indicative features or otherwise selecting high-level model parameters.
In the IML process the user becomes the principle driver of the interaction to provide the desired behaviour in the system. Giving the user the ability to operate on the system does not imply that the computer has no influence on the process or

does not make independent decisions. Indeed, the application may, for example, intelligently select a subset of data for review. The concept of the user as the main driver of the operation is seen in the fact that the IML application seeks to provide the user with control over the high level behaviour of the system. The user drives the process by providing feedback and model training. The interface is the bridge between the user and the model and data and provides the basis for interaction. The interface component is the primary focus of this Chapter.
The user may possess significant domain expertise relevant to interpretation of the data and evaluation of model outputs. In [61] three important aspects of an IML process are highlighted that can be thought of as desired interface attributes: 1) illustrate the current state of the concept learned; 2) guide the user to provide inputs that improve the quality of the concept; and 3) provide review mechanisms that allow the user to explore the model space. [62] identify three activities relevant to the productive integration of the human user and machine learning techniques: 1) transmitting the reasoning of the system to the user; 2) transmit the user's reasoning to the system; and 3) ensure both the system and the user profit from this feedback loop. The four key elements of the interface are therefore: sample review, feedback assignment, model inspection and activity overview [63].
A decisive aspect of the growth of human-computer systems is Usability. Usability issues are still identified late in the software development process, during testing and deployment. One of the reasons for the delay in identifying these issues is in the system design during requirements assessment which does not incorporate usability perspectives effectively into software requirements specifications. The main strength of usability-focused software requirements is the clear visibility of usability aspects for both developers and testers.
Design science is described as an inventive or creative problem solving activity [64], and focuses on how to develop and produce artefacts and artificial systems having desired properties. In [65] the importance of design activities for the discipline of information systems (IS) is emphasized and presenting a conceptual framework for understanding, executing, and evaluating IS research combining behavioural science and design science paradigms. An important aim of Human Computer Interaction is to gain a detailed understanding of cognitive, perceptual, and motor components of user interactions with human computer systems [66]. The two models to be used to develop a detailed understanding and to elaborate the functional specification are the user and usability models. These models will be formally designed as specifications [67].
A human-centered understanding of machine learning in human context can lead not only to more usable machine learning tools, but to new ways of framing learning computationally. [68] studied expert programmers working with machine learning and identified a number of difficulties, including treating the methods as a "black box" and the difficulty of interpreting the results. A human-centered approach to

machine learning that rethinks algorithms and interfaces to algorithms in terms of human goals, contexts, and ways of working can make machine learning more useful and usable. In interactive machine learning, the user chooses what new examples to label and/or create, or works together with the algorithm in controlling the process. In this way, the computer is part of a human design process, rather than the human being in the loop of an algorithmic process.
Past work also demonstrates ways in which a human-centered perspective leads to different approaches to evaluating, analysing, and understanding machine learning methods [20]. Design research suggests that explicit mechanisms to support exploration, comparison of alternative prototypes, and iterative refinement are fundamentally important to enabling efficient and effective design [69, 70]. Machine learning tools should explicitly aid users in these activities. Tools must also provide effective feedback to inform subsequent user actions: to help users understand how to debug a model that has not learned a concept correctly, to understand the trade-offs between different formulations of a learning problem, and even to understand the limits of what can be learned [71].
Early computer software aimed to solve business and scientific problems in a predetermined way that allowed only very constrained user input, through arguments given to the program at runtime. This contrasts sharply with modern day software, which is much more interactive and supports frequent user input as it runs. This shift towards interactive software is reflected in the growing emphasis on interfaces designed to facilitate communication between software and humans.
However, one major drawback of existing interactive systems is that they have little ability to take into account differences in the knowledge, style, and preferences of their users. Clearly, there is a need for increased personalization in many areas of interactive software, not only in the types of flexibility but in the way that personalization occurs. Moreover, some facets of user styles may be reflected in their behavior but not subject to conscious inspection. This suggests the use of techniques from machine learning to personalize interfaces, based on the observation of user activity. An adaptive user interface is an interactive software system that improves its ability to interact with a user based on partial experience with that user. Rather than replacing a human, the system suggests information or generates actions that the user can always override. Ideally, the learned knowledge should reflect the preferences of individual users, thus providing personalized services for each one.
However, this focus on advisory systems leads directly to another characteristic: the user's decisions give a ready source of training data to support learning. Every time the interface suggests some choice, the human either accepts that recommendation or rejects it, whether this feedback is explicit or simply reflected in the user's behavior. Either way, the system obtains another datum to drive its search for an improved knowledge base, and each case includes details about the decision-making

situation, providing important context for future predictions.
The embedded nature of the induction process has another implication for the learning task: the system should carry out online learning, in which the knowledge base is updated each time an interaction with the interface occurs. This contrasts with most work in data mining, which assumes that all data are available at the outset. Because adaptive user interfaces collect data during their interaction with humans, one naturally expects them to improve during that use, making them 'learning' systems rather than 'learned' systems. Because adaptive user interfaces[72] must learn from observing their user's behavior, another distinguishing characteristic is their need for rapid learning. Still, adaptive interfaces that learn rapidly will be more competitive, in the user's eyes, than ones that learn slowly.
There are two types of adaptive user interfaces: informative interfaces and generative interfaces. Informative interfaces attempt to select or filter information for the user, presenting only those items he will find interesting or useful. Generative interfaces, focuses on the generation of some useful knowledge structure [72].
The goal of visual analysis (VA) systems is to solve complex problems by integrating automated methods of data analysis, such as machine learning (ML) algorithms, with visualizations. More importantly, it is crucial to incorporate the knowledge, intuition and feedback of the human being into the analytic process, so that hypotheses can be refined and models can be tuned. By integrating ML algorithms with interactive visualizations, VA aims to provide a visual platform for the analyst to interact with their data and models [73]. Interactive visualizations act as an aid or "lens" that facilitates the process of interpretation and validation, but they also make ML interactions accessible to analysts. Usually, simple exploration interactions, such as changing visual coding or navigation, don't feed back into the ML components.
In VA systems, analysts are actively involved in an iterative process of observing, interpreting, and validating system results followed by subsequent refinement. Such an approach would favor the direct use of ML tools by domain experts. Visual interfaces that are easy-to-use and understand allow such analysts to introduce their domain knowledge more effectively and consequently adapt ML components in order to further progress in data-intensive but ill-defined analysis tasks[74] .
Overall, VA tools have the potential to enhance the support of interpretation, understanding, validation and refinement of ML through interaction. However, current VA tools and ML components are posing many interesting challenges for future work. To address these challenges, closer collaboration between ML and visualization researchers is vital [75].
UX designers today have difficulty in interfacing with ML systems because they do not have suitable prototyping tools to work with ML. It is difficult to quickly prototype and understand the user experience impact of false negative and false positive responses from an ML service. Finally, it could be that UX designers don't

have a clear understanding of what ML is and what it can do.
Recent UX articles on the web, where UX designers talk about ML, often reveal huge misconceptions about what ML can actually do, with many designers treating it too much like magic. Research on human-robot interaction shows that physical proximity, organizational state and structure of activities can alter people's experiences [76], and that setting expectation and recovery strategies help mitigate errors [77]. The value of design comes from instilling particular products and services with the quality of experience that distinguishes them from everyday life; and without good tools, designers have difficult to explore the space of possibilities.
A survey conducted by [78] revealed how much ML is seen as something that is just starting to be important and will be more important in the future. It shows that designers don't have a clear understanding of ML technology and how to imagine uses that they don't yet they exist and possibly explain why UX designers haven't fine-tuned their expertise in using ML in today's commercial products. Respondents noted the need for designers to collaborate with skilled technicians noting that machine learning is difficult to prototype. ML clearly requires a new type of prototyping, which does not yet exist.
Machine learning implies that the system and data will change over time, and designers are not used to designing a form for large-scale dynamic data. Also MLs and designers do not treat data in the same way. Designers primarily visualize data and look for meaningful correlations and patterns, which fit their understanding of how the world should and works. On the contrary, Ml finds machine-recognizable correlations and patterns in the data.
The survey results[78] showed that ML is considered technically complex and challenging. The interviewees described the difficulties of understanding and therefore expressing the capabilities, limitations and potential of ML within a UX design context. The statistical intelligence shown by ML can lead to a very different interpretation of the same data than the human intelligence of common sense. This can make performance errors bizarre and difficult to explain, resulting in potentially dissonant user experiences.
To mitigate this, UX designers should consider interactions not only from the more familiar human perspective, but also from the perspective of statistical inference and, fundamentally, how these two perspectives might interact. The challenge for researchers is not only to undertake such a review, but also to develop the tools that allow us to explore the results with potential users.
Machine learning (ML) is now a fairly established technology, and user experience (UX) designers have begun to integrate ML services into the things they design.

5.1 Interactive Notebooks

In the previous introduction, the importance of greater user involvement in the analysis process was emphasized. This extra aspect to take into consideration can create a certain annoyance, in the absence of a User Experience design expert, for the data scientist, if he has to take care of the interactive aspect that is not his responsibility. The design of a user interface, of the workflow, of an easily usable interaction, can be very time-consuming and it is for this reason that it is often put aside. In this section we want to show how user involvement can also occur at a lower level, not requiring any additional time compared to the analysis that is taking place, but integrating a certain degree of interaction within the code itself, exploiting the potential of existing tools that are having a wide diffusion among data scientists.
The tool we're talking about is Notebook Jupyter [79], a free, open-source interactive web tool known as a computational notebook that researchers can use to combine software code, computational output, explanatory text, and multimedia resources into a single document. From a user perspective, notebooks Jupyter offers a convenient web-based user interface for iterative code execution, output exploration, and data visualization, all from a single environment [80, 81, 82].
Michael Bostock [1][83] defined the notebook as an interactive and editable document defined by code, a computer program, but designed to be easier for humans to read and write. A Jupyter notebook consists of two key modules: a user interface and a kernel. The user interface is where you edit your notebook by adding cells that can contain text, code, images, or other elements like maps and charts, writing code, and explaining the results. It is a web application that runs in your browser. The kernel, in our case for Python, is where the code is executed. It is a separate process that is done outside of the browser. When you execute a cell of code, the kernel calculates the result and sends it back to the browser, where you can continue working on it. This separation between the user interface and the kernel makes Jupyter notebooks highly modular.
A problem that can be encountered in the use of the Jupyter notebook is the need to have to rerun one or more cells several times when changing input parameters, making the analysis process not only inefficient but also frustrating, interrupting the flow of an exploratory data analysis. To overcome this problem there is a tool to build interactive controls in Jupyter that allow you to modify the inputs without the need to rewrite or re-execute the code, the IPython widgets [84].
The ultimate goal is to be able to create an environment as interactive as possible

[1]Michael Bostock is a famous computer American scientist, specialist in data visualization, is one of the co-creators of Observable and is known as one of the key developers of D3.js.

while remaining in the context of the notebook. This type of approach is less effective if we want to increase the diffusion of the work because it limits the distribution to that circle of people capable of understanding and manipulating the code. However, it remains a valid alternative to improve the user experience and interest the recipients of the analysis. The following will describe how this type of approach for the analysis described in Chapter 4 was implemented. The idea behind this step is to make the user author of the analysis, allowing as much as possible the choice of the parameters used and showing easy to understand interactive results.

5.1.1 Selection of the dataset

The first operation requires the user to choose the dataset. The user can select one of the datasets provided by clicking on one of the toggle buttons shown in the figure and confirming to save the changes, or he can load a dataset from his file system (the available datasets are already discretized, any dataset loaded by the user must be already discretized, a control verifies this condition).

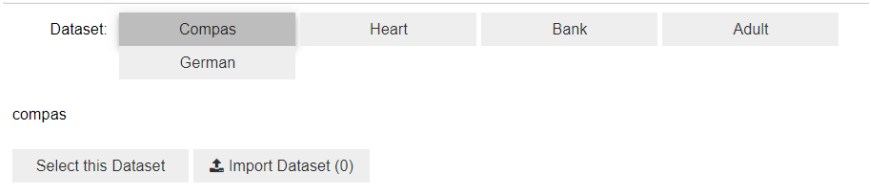
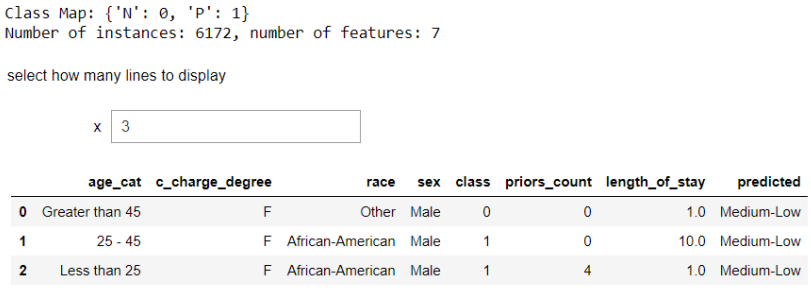


Figure 5.1: Display Dataset selection

Once the selection has been made, if a dataset has been chosen from those available, the user will see the class map of the dataset, the number of instances, the number of features, and the first three rows of the pandas Dataframe [85] in table form, the user can decide to display more lines by changing the number of the text box. In our running example, we select the COMPAS dataset [54].



	age_cat	c_charge_degree	race	sex	class	priors_count	length_of_stay	predicted
0	Greater than 45	F	Other	Male	0	0	1.0	Medium-Low
1	25 - 45	F	African-American	Male	1	0	10.0	Medium-Low
2	Less than 25	F	African-American	Male	1	4	1.0	Medium-Low

Figure 5.2: Selected Dataset's information

If instead, the user has decided to import a dataset of his choice, it is necessary to choose the features that will represent the class map. In this case, a select is provided to decide the attribute that will be used, once the field has been chosen, the user must select the value that will represent the positive class, click on the confirm button and display a preview of the dataset consisting of the first five lines.

Figure 5.3: Selection of the class map for dataset uploaded by the user

5.1.2 Analysis parameter set and Evaluation metrics selection

Once the dataset has been selected, the user must set the parameters necessary for the analysis. A top box allows you to choose the support threshold value, two other text boxes allow you to indicate the true class and the predicted class, once confirmed the user can choose the evaluation metrics of his interest through multiple selectors with all available metrics (if the aforementioned class is not indicated, the False Positive Rate ("FPR"), and the False Negative Rate (" FNR ") will be available as evaluation metrics).

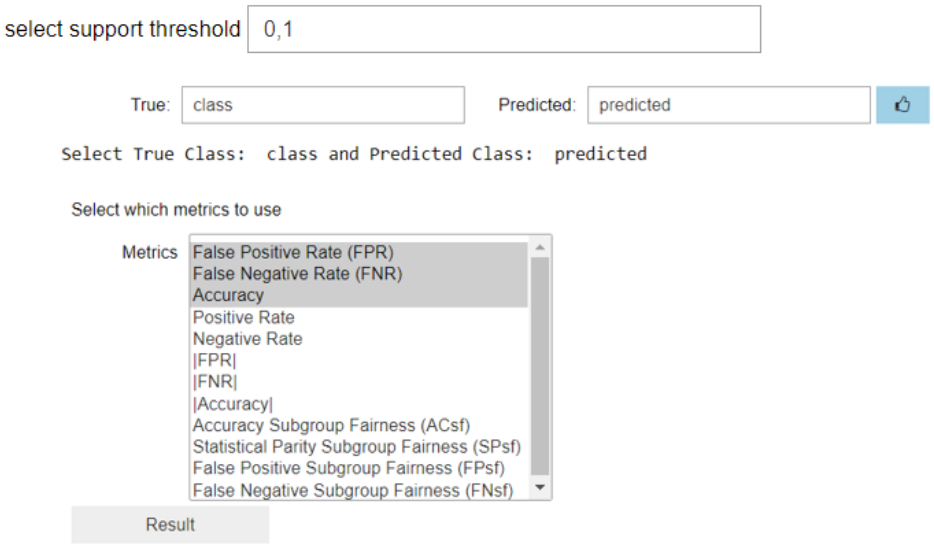
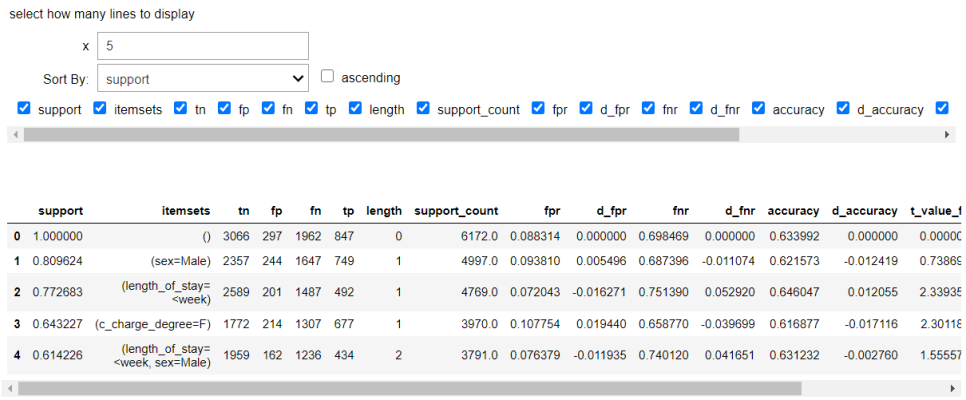


Figure 5.4: set of analysis parameters

5.1.3 Frequent patterns extraction

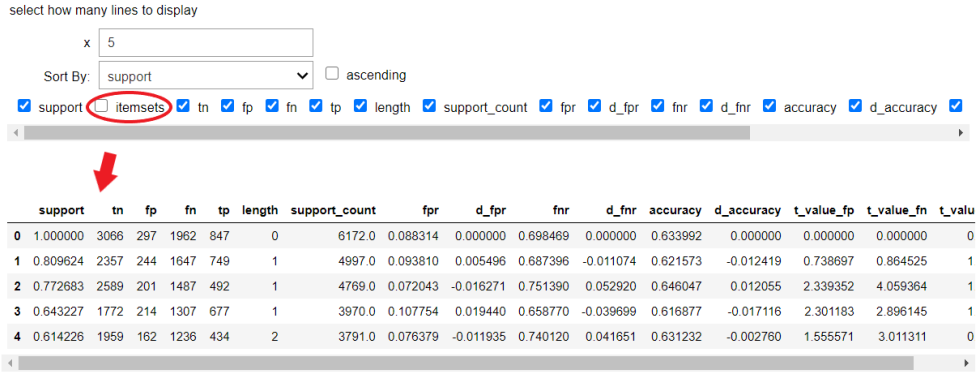
From this point begins the analysis, built to accompany the user throughout the process. The first phase consists of a display of frequent patterns, sorted according to the supporting value in which the user can evaluate, for each set, the calculated value for both the classification metrics and the discrepancy metrics selected previously. To facilitate the evaluation, the user has a series of commands including the selection of the number of lines, the selection of the field to be used in sorting, and a checkbox to change the sorting order (ascending/descending).



	support	itemsets	tn	fp	fn	tp	length	support_count	fpr	d_fpr	fnr	d_fnr	accuracy	d_accuracy	t_value_f
0	1.000000	()	3066	297	1962	847	0	6172.0	0.088314	0.000000	0.698469	0.000000	0.633992	0.000000	0.00000
1	0.809624	(sex=Male)	2357	244	1647	749	1	4997.0	0.093810	0.005496	0.687396	-0.011074	0.621573	-0.012419	0.73869
2	0.772683	(length_of_stay=<week)	2589	201	1487	492	1	4769.0	0.072043	-0.016271	0.751390	0.052920	0.646047	0.012055	2.33935
3	0.643227	(c_charge_degree=F)	1772	214	1307	677	1	3970.0	0.107754	0.019440	0.658770	-0.039699	0.616877	-0.017116	2.30118
4	0.614226	(length_of_stay=<week, sex=Male)	1959	162	1236	434	2	3791.0	0.076379	-0.011935	0.740120	0.041651	0.631232	-0.002760	1.55557

Figure 5.5: display of frequent patterns

Furthermore, the user can decide to eliminate unwanted columns by removing the check of the column in question, as shown in the Figure 5.6.

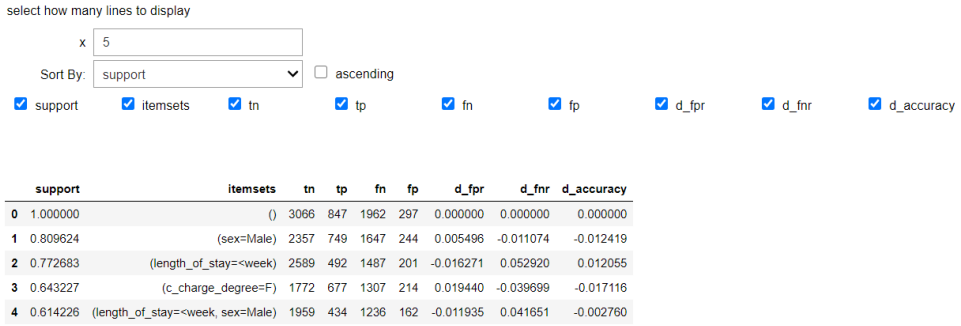


	support	tn	fp	fn	tp	length	support_count	fpr	d_fpr	fnr	d_fnr	accuracy	d_accuracy	t_value_fp	t_value_fn	t_valu
0	1.000000	3066	297	1962	847	0	6172.0	0.088314	0.000000	0.698469	0.000000	0.633992	0.000000	0.000000	0.000000	0
1	0.809624	2357	244	1647	749	1	4997.0	0.093810	0.005496	0.687396	-0.011074	0.621573	-0.012419	0.738697	0.864525	1
2	0.772683	2589	201	1487	492	1	4769.0	0.072043	-0.016271	0.751390	0.052920	0.646047	0.012055	2.339352	4.059364	1
3	0.643227	1772	214	1307	677	1	3970.0	0.107754	0.019440	0.658770	-0.039699	0.616877	-0.017116	2.301183	2.896145	1
4	0.614226	1959	162	1236	434	2	3791.0	0.076379	-0.011935	0.740120	0.041651	0.631232	-0.002760	1.555571	3.011311	0

Figure 5.6: display of frequent patterns with dropped column

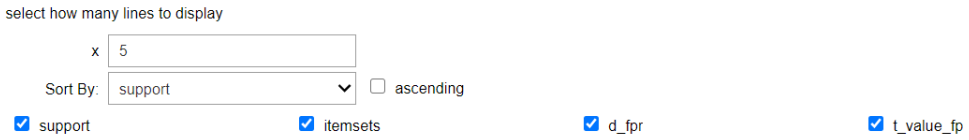
This type of representation will be presented to the user whenever you want to give the possibility to specifically evaluate the values obtained by allowing the sorting of the rows and the drop of the columns. We find, in fact, this type of scheme even later, when the results obtained for the selected metrics are individually evaluated. Let's take the example of the False Positive Rate (FPR).
The first operation that the user can do is the evaluation of the most discrepant patterns, comparing for each pattern the values of the discrepant metrics or even just the metric of their interest, eliminating the columns of no interest or displaying immediately below the table with only discrepant information, being able to select the number of rows to display, the field used for sorting, if you want an ascending/descending sorting and being able to eliminate the columns that you do not

want to take into consideration.



	support	itemsets	tn	tp	fn	fp	d_fpr	d_fnr	d_accuracy
0	1.000000	()	3066	847	1962	297	0.000000	0.000000	0.000000
1	0.809624	(sex=Male)	2357	749	1647	244	0.005496	-0.011074	-0.012419
2	0.772683	(length_of_stay=<week)	2589	492	1487	201	-0.016271	0.052920	0.012055
3	0.643227	(c_charge_degree=F)	1772	677	1307	214	0.019440	-0.039699	-0.017116
4	0.614226	(length_of_stay=<week, sex=Male)	1959	434	1236	162	-0.011935	0.041651	-0.002760

Figure 5.7: Most discrepant patterns with all selected metrics



	support	itemsets	d_fpr	t_value_fp
0	1.000000	()	0.000000	0.000000
1	0.809624	(sex=Male)	0.005496	0.738697
2	0.772683	(length_of_stay=<week)	-0.016271	2.339352
3	0.643227	(c_charge_degree=F)	0.019440	2.301183
4	0.614226	(length_of_stay=<week, sex=Male)	-0.011935	1.555571

Figure 5.8: Most discrepant patterns with only discrepant information

The interaction remains the main point of the construction of the notebook, but this thesis also wants to demonstrate how a different representation of the data can increase the user's interest in what he is doing.
We show below the next step of the analysis, the calculation of the top-K patterns, in which the user cannot manipulate the calculation but only provide the K parameter to decide how many values to observe. In this case, we have chosen to change the standard representation and to build a special table with the plotly library [86], thanks to which it was possible to set the size of the table, the size of the individual columns as well as change the color by enhancing the output obtained which will be used then later as a basis for some choices.

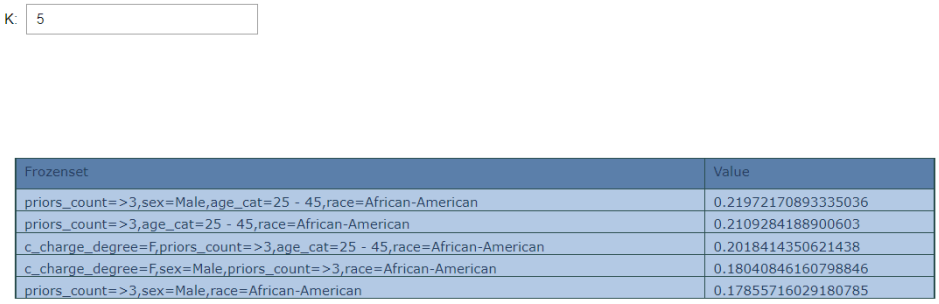


Figure 5.9: Top K table with selection of the number of rows

Once the table has been obtained, the user can view the Shapley value, represented as a bar graph, by selecting the group corresponding to the row of the above table whose output is to be evaluated. The group value can be selected through the use of the toggle widget that allows you to create clickable buttons associated with the plot corresponding to the number indicated above. Furthermore, the realization of the bar graph through the plotly library allows to obtain an interactive graph, in fact, when the mouse passes over a bar, a tooltip will appear, that is a small help window, containing the information relating to that bar.

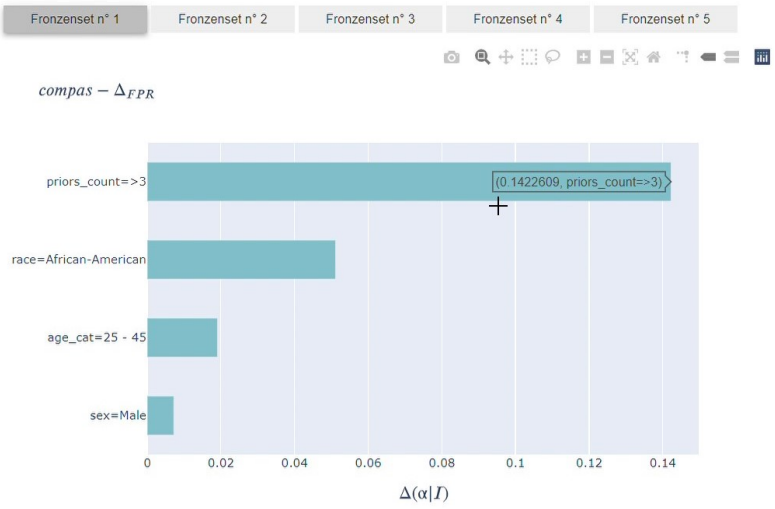


Figure 5.10: Bar chart for Shapley Value

Subsequently, the choice of the pattern will also be presented in the evaluation of the lattice graph, in which the user can set the group discrepancy threshold. The visualization of lattice search shows a lattice starting from a root with an empty pattern, building the second level with nodes containing only one element per pattern, and so on going down. Furthermore, the value of the score for that pattern is shown, diversified through the use of 3 colors. Red indicates a score value greater than the group discrepancy threshold, dark blue indicates a lower group discrepancy value, the light blue square indicates a lower score value, in the descent than the previous one.

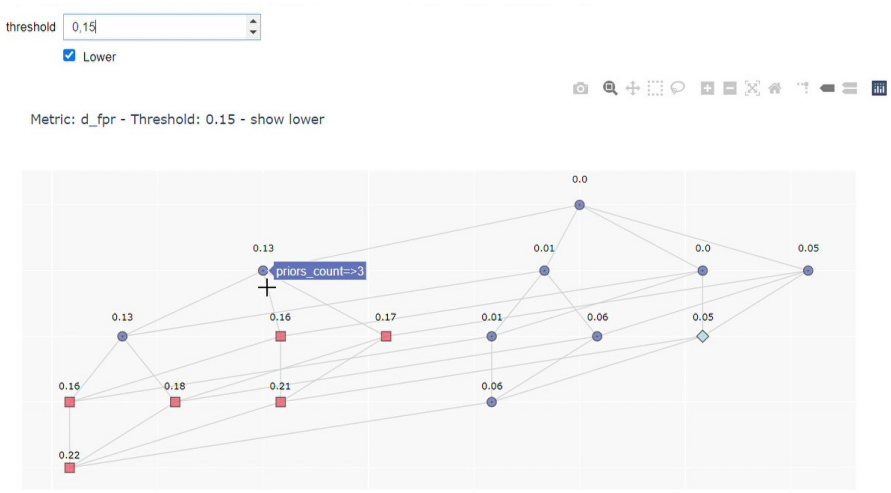


Figure 5.11: Example of Lattice Search visualization

By clicking on the appropriate "lower" checkbox, the user can decide not to show when the score has decreased in the descent, but will only display the corresponding symbol relating to the score value (blue circle for values below the set threshold, red square for values exceeding). We also note here that the use of the plotly library allows us to build an interactive graph that shows information about the node on mouseover.
The objective of the analysis is the search for problematic subgroups that can lead the classifier to an incorrect prediction, to do this we must look for those subgroups that influence the behavior of the classifier.
The next step of the analysis involves the evaluation of the contribution, positive or negative, which causes the addition of an item to an itemset.
In this case, the user can view the summary table with the first five values, where at the user's choice it is possible to decide to increase or decrease the number of lines displayed by simply changing the number in the text box immediately above, the display update will update automatically.

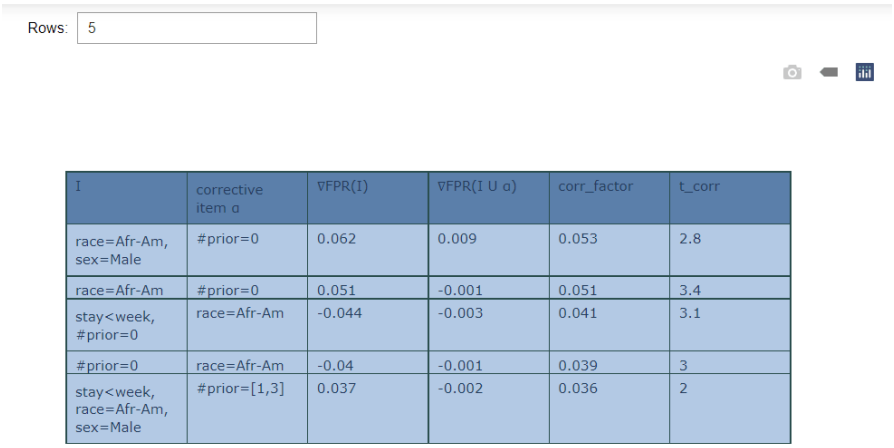


Figure 5.12: Table with Corrective values

From the vision of the table, we move on to the comparison between the Shapley Value of the original item and the Shapley Value of the item set with the addition of the corrective item. With reference to the table, the user can select the item to be taken into consideration and view the differences between the two bar graphs.

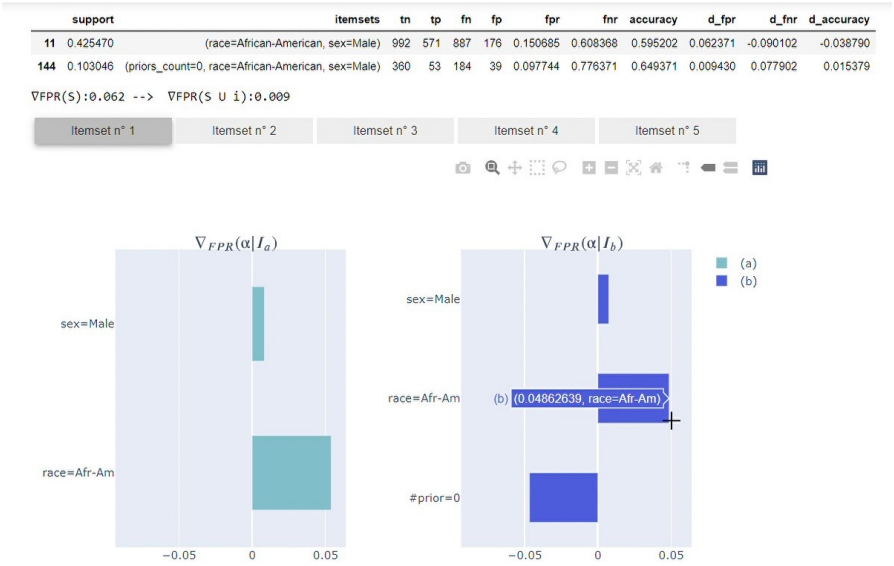


Figure 5.13: Comparison of the Shapley Values of two subgroups

Having evaluated the Shapley value, the user can also evaluate the lattice search of the new itemset with the corrective item. In this way he will be able to understand what the contribution of each subgroup is in the final prediction. It can modify the group discrepancy threshold and decide whether or not to show differently, the subgroups that have a lower score in the descent of the lattice.

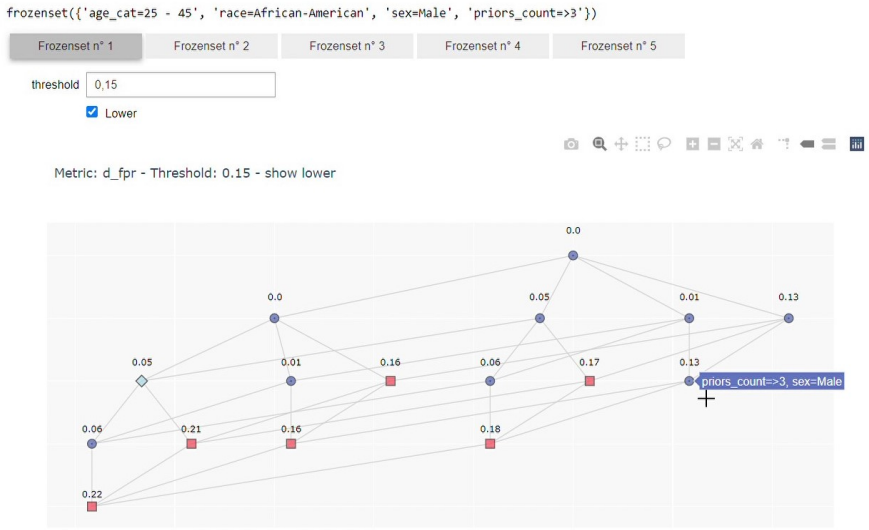
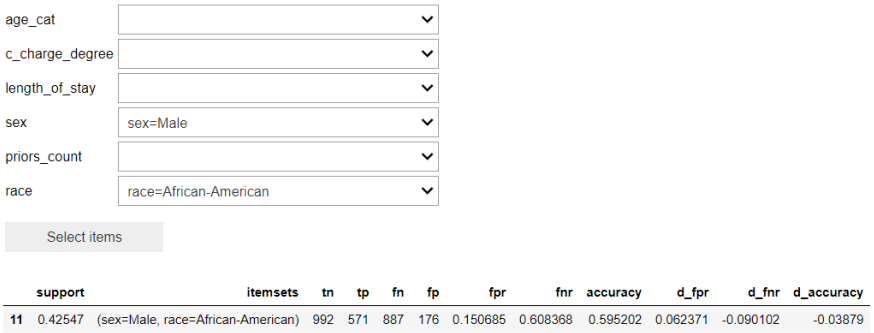


Figure 5.14: Visualization of lattice search of itemset with corrective item

The last aspect that the user can evaluate of the single metric taken into consideration is the single subgroup chosen by the user himself. For each field of the dataset, the user has a selection with all possible values for that field. By choosing, for example, "sex = Male" and "race = African-American" and confirming with the "Select items" key, we obtain the single row of the dataset with all the calculated values.



	support	itemsets	tn	tp	fn	fp	fpr	fnr	accuracy	d_fpr	d_fnr	d_accuracy
11	0.42547	(sex=Male, race=African-American)	992	571	887	176	0.150685	0.608368	0.595202	0.062371	-0.090102	-0.03879

Figure 5.15: Selection of values for each field with display of the selected item

If the combination of the chosen values presents a support value below the threshold set at the beginning, the user will be informed, who can decide whether to display the same result or reformulate another choice, as shown below.

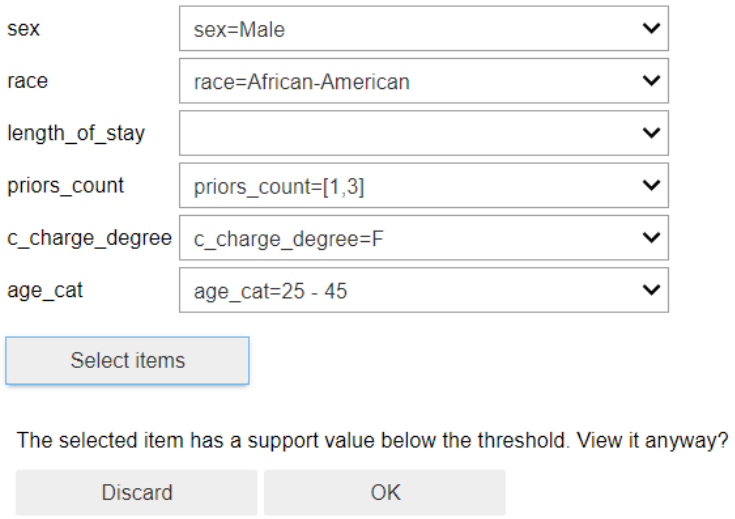


Figure 5.16: notice to the user item below threshold

All the operations described above are repeated for each metric selected at the start of the analysis. Finally, to conclude the analysis, we provide the user with a summary representation of the results obtained for each metric, showing both the overall result and a comparison plot, so that the user can have a clear view of what he has just analyzed, can see the trends of the various contributions and, if necessary, with a simple mouse hover, observe the corresponding numerical value of the Shapley.

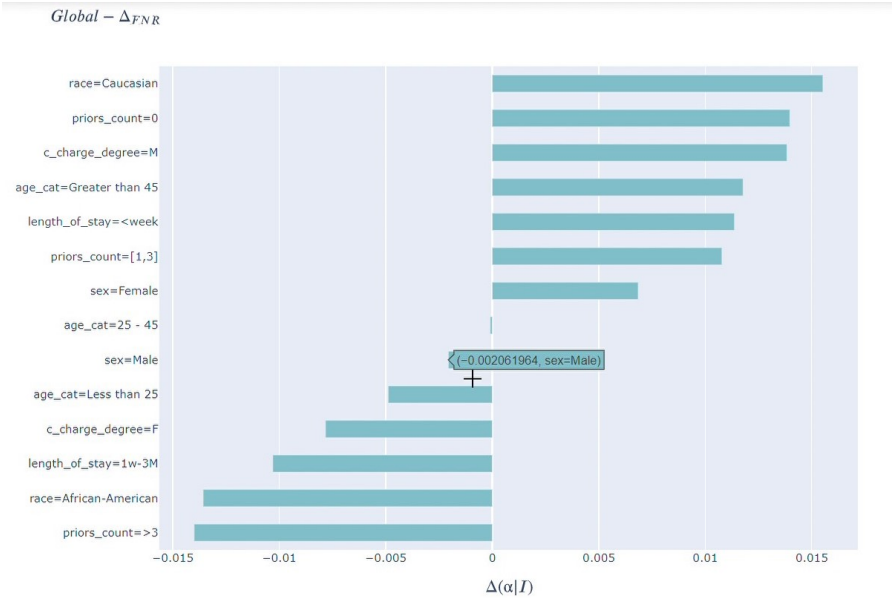


Figure 5.17: Global Shapley Value

Figure 5.18: Global Comparison Shapley Value

Designing a notebook in this way has a double benefit. On the one hand, it facilitates the comprehension process through the abundant use of graphic representations which, as we know, constitute a mechanism for transferring information that the human brain prefers over text and helps to simplify and categorize the fundamental operations carried out, on the other hand, the more experienced user can also evaluate the implementation details and experiment with new hypotheses through the interactions, manipulating the modifiable parameters.

5.2 User Interface

In this section, we will deal with the description of the user interface for the interactive tool that has been created. At the base of the design of the user interface there is the concept of usability, defined by the International Organization for Standardization [87] as the effectiveness, efficiency, and satisfaction with which certain users achieve certain objectives in certain contexts, i.e. it indicates how easy and satisfying to use a tool when the user interacts with it. Usability is not a characteristic of the system but must be understood as a property resulting from human-computer interaction, we can see it as a measure of the cognitive distance between the "design model", the product model and its functioning conceived by the designer, and the “user model”, the idea created by the user of what the product should look like and how it works. The closer the two models are, the less usability is a problem [88]. This demonstrates how important a good and efficient design of the user interface is, as it must be clear to the user, from the first use, the possibilities, limits, and operating modes with which he will interface, without difficulty he must understand the actions that are possible on the interface, and mostly what results he can achieve.
The quality of software can be evaluated by considering three aspects: the first aspect concerns the functionality, the second the usability, and finally the third the user experience. To ensure the usability of the final product it is important to try to satisfy some requirements [89]. First of all, ease of learning must be guaranteed, the user must be able to achieve good performance in a short time; the system must be efficient in terms of performance; ease of remembering, the user must be able to interact with the interface even after a long period of inactivity without having to start from scratch; the system must have a low probability of error and must be robust in case of error; finally, it must guarantee a certain degree of satisfaction during the interaction.

5.2.1 Front-end Environment

The front-end part of the tool was developed using ReactJS[90], a JavaScript library widely used in the creation of modular user interfaces. The use of ReactJS allows the development of large web applications that can modify data without subsequent page updates, and is widely used due to its highly efficient execution [91]. Considering the classic MVC (Model-View-Controller) design paradigm used in user interface development, ReactJS positions itself as View. An efficient and light virtual DOM [92] (Document Object Model) is created saved in memory with which React interacts without touching the DOM generated by the browser, leading to fast and above all robust performance of the application, because before reflecting the changes of a page on the Web, React makes the changes to the virtual

DOM. After modifying its copy in memory, React applies a comparison algorithm between the two DOMs and updates only the desired nodes of the browser DOM, thus avoiding rendering the entire DOM. The DOM represents the physical display, created according to an HTML model, of the tree of the components that make up the user interface.

5.2.2 Back-end Environment

The Back-end was developed using Flask [93], a Python micro framework that provides the core functionality of the web framework and allows you to add more plug-ins so that the functionality and feature set can be extended to a new level. A framework is a library or collection of libraries that aims to solve a part of a generic problem instead of a completely specific one. Flask is defined as micro because it makes the basic functionality simple but extensible in terms of development, that is, it implements only basic functionality (including routing) but leaves more advanced features (including authentication and database ORM) to the extensions, among the services offered we find the integrated HTTP server, support for unit tests and the RESTful web service. This makes writing applications or extensions very easy and flexible and gives developers the power to choose the configurations they want for their application, without imposing any restrictions on the choice of database, model engine, and so on. The result is less initial setup for the first-time user and more choice and flexibility for the experienced user. Flask uses Jinja Template Engine [94] and Werkzeug WSGI Toolkit [95]. Flask structure is divided into two parts, Static files and Template files, the template file has all Jinja templates including Html pages, whereas a static file they have all the static codes needed for the website such as CSS code, JavaScript code and Image files. Once imported into Python, Flask can be used to save time building web applications. In the tool described below, Flask routing is used a lot with the structure in Figure 5.19

```
from flask import Flask
app = Flask(__name__)

@app.route("/")
def hello():
    return "Hello World!"
```

Figure 5.19: Flask URL routing in Hello Word web application

as we can see, Flask uses decorators for URL routing, that is, it assigns a certain function to the route we indicate (in this case defined by the single bar), it is a Python shortcut that allows you to call the function indicated under the decorator every time the user visits the page of our web application defined by the indicated route. In reality, this system can also be used in the absence of the page corresponding to the route when the webserver receives an HTTP request from the web browser to the specified route, it can also be useful for passing values from the browser to the server. The web server, in case of a GET request, will reply, in our case, with a JSON object which can be described as a key/value dictionary. For more information regarding Flask see the following references [96, 97, 98, 99, 100].

5.2.3 Graphical User Interface (GUI)

Up to this point we have talked about how important it is, for a good diffusion, to design a user interface to facilitate the understanding process. In this section, we will describe the user interface in the composition of its graphics and in the stylistic choices that have been made as it is not sufficient to present a set of basic elements positioned, randomly, on a blank page. The composition of the graphic elements and their organization on the page is the first aspect to consider to improve the user experience. We find this concept well expressed in the philosophy of Steve Jobs [101, 102] who put the user experience at the center of his projects, also researching the particular design choices that would have involved the user more from a sensorial point of view. Steve Jobs abstracted the concept of design to a higher level, it is not only aesthetic but represents the entire functioning of the product, as it is the relevant factor that determines its success or failure because it represents how the product will be perceived and used by the user.
The user interface that has been designed has the purpose of facilitating the learning process, for this reason, we can say that it has an educational purpose [103]. An important aspect to consider when designing educational software is the time that the user will spend using it, it will not be very long but limited to the time needed to learn the software content. This consideration is the basis from which to start in the design of the interface, it must not require too much effort on the part of the user to understand its functioning which does not represent the ultimate goal to which it is aimed, adding self-describing and indicative parts to the design that facilitates an automatic use of the tool. Making a similarity, we can say that the parts of the description represent a bit of the road signs of the path that we want to make the user do, indicating the right way to go. To make the interaction more efficient from the first use, the interface is built so as not to project the user directly into the analysis but presents a sort of pre-phase of the general presentation of the tool in a congenial scheme that many applications they use, i.e. home page (Figure 5.20, about page (Figure 5.21), and start page (Figure 5.22).

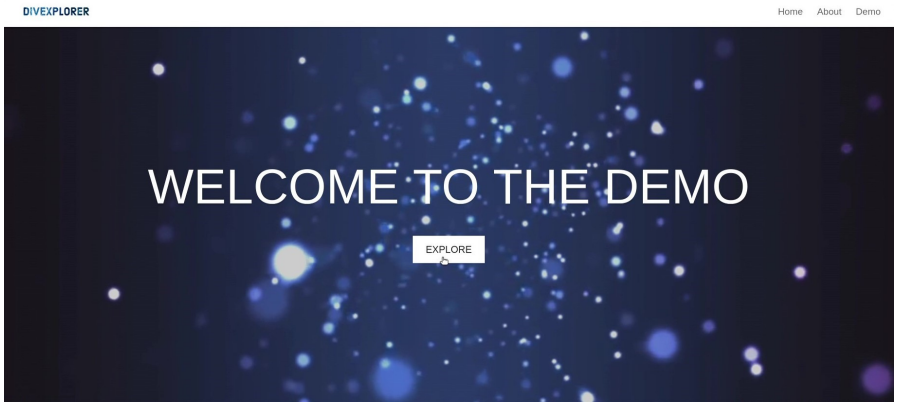


Figure 5.20: First page of proposed Interactive Tool

The setting of anticipation of the key points of the analysis is what makes the presented system effective, the user is quickly guided to assimilate the key points of what he will actually find himself exploring, the About page has precisely the purpose to provide the user with a summary explanation of the problem and the proposed solution, furthermore, in addition to the textual explanation there is a carousel which, with a simple scrolling animation, shows the user some details of the tool (Figure 5.21).

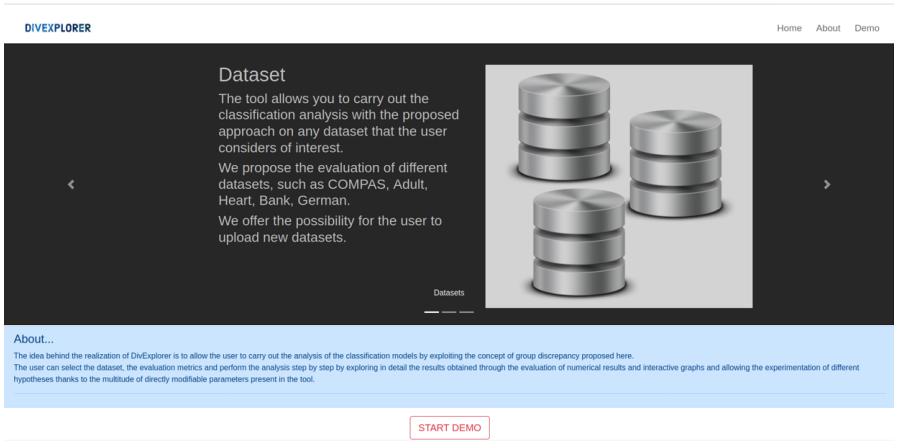


Figure 5.21: About page of proposed Interactive Tool: the page is composed at the top of the navigation bar, in the center we find an automatic scrolling carousel that shows some details of the process, under a text bar containing a small summary of the problem and the proposed solution.

If the user decides to skip the About part and wants to start the analysis directly, the tool stops the user on an intermediate page, since even if he has not read the description of the process that will be carried out, it is necessary to provide the user with some instructions to follow. The basic idea is to guide the user during each step ensuring smooth execution of the actions, the user usually tends to be hasty in the preparatory phase, skipping the reading phases to immerse himself in the active, interaction phases. To mitigate the consequences of the tendency to skip important steps that could lead to a non-sliding demo, it is good practice to insert additional actions to be presented to the user, which could be a confirmation or data entry. In this case, it is simple feedback in the form of a button, where the user confirms that he wants to start the analysis, and even if he has not read the suggested instructions, during the demo the user will receive alerts relating to the case of possible error specific to the action he is taking.

Figure 5.22: Demo page of proposed Interactive Tool: this page is an extra step proposed to the user before starting the exploration to underline some important instructions that must be taken into consideration during the analysis

The dynamism of an interface is what makes it more appealing to the user and from the first page we wanted to pay attention to details. The first aspect taken into consideration was the selection of the palette to be used in the pages because thanks to the colors it is possible to express what you want to communicate, it is possible to emphasize a certain action, and it is the first detail that involves the user as the colors generate emotions and have meanings. It is a good idea not to make too much use of colors and strong contrasts with highly saturated colors because they can annoy the user's eye [104]. In our case we have chosen to use the classic colors of the technology sector, the main background of each page of a

non-bright white that allows the right contrast with the objects in the foreground. The elements inside the page have been created using gray that creates a good but not excessive contrast. Finally, for the other details, we have chosen to use a variant of blue because as Boyle [105] states *"the distribution of cones in the eyes also makes peripheral vision of blue over large areas quite effective"* and also, it is a color that conveys tranquility and makes more productive.The possible resolvable errors are orange which attracts the user's attention and finally the success of the operations is indicated in green, classic color to indicate positivity.

Figure 5.23: Display examples of indications: on the left an example of a warning, to indicate to the user that he is carrying out a wrong action; on the right an example of successful action, the user is reported of the success of an operation

Another aspect taken into consideration concerns drawing the user's attention when she is about to take a step that involves changes relevant to the result. Attracting attention means stimulating the user's perception that something has changed and we do it by exploiting the properties of the objects that make up the page. For example, the user can accidentally confirm some settings she did not want, or start a wrong count because distracted from the moment, she did not pay the right attention to where he clicked, this is a fairly common situation. The Figure 5.24 shows an example of how to avoid any distractions simply by making the confirmation action dynamic, a button with only the marked outline that fills up when the mouse passes over it creating contrast with the background can be a good solution. The user will be captured by the element that at that moment takes on a different aspect, standing out on the page, attracting her attention, and thus avoiding involuntary actions.

Figure 5.24: Dynamic button example display: on the left the layout of the button when it does not receive interaction; on the right the button when the mouse is positioned over it.

Finally, we present the general scheme that has been adopted for all pages from the moment the user begins the analysis.

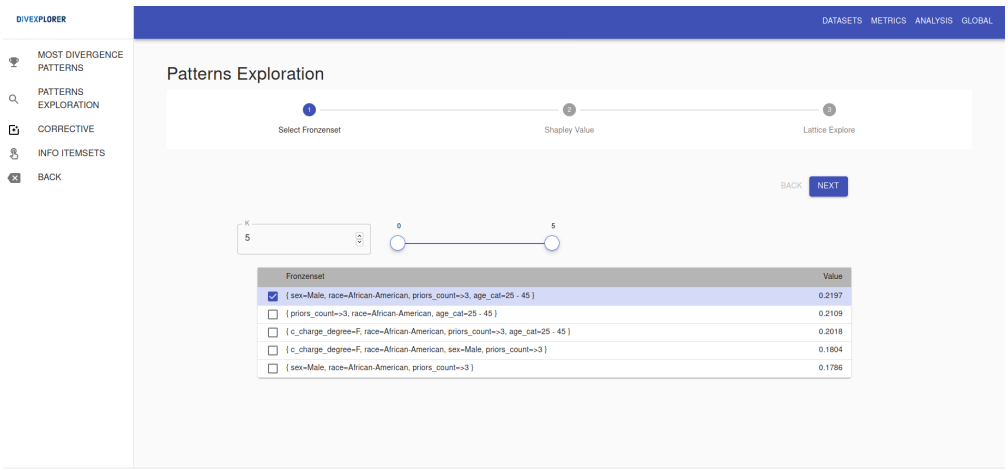


Figure 5.25: Example of page: all the pages of the tool have this scheme, navigation bar at the top for the main steps, side navigation bar for the single operations relating to the step, central content with the operation to be carried out.

From the Figure 5.25, we can see the presence of all the elements necessary to guarantee the user an effective and smooth interaction. The user must always be able to understand where she is navigating. For this reason, there are two navigation bars, at the top, there is the navigation bar of the macro elements of the analysis plus the logo that allows the return to the home page. It is the main navigation bar that summarizes the key points of the exploration allowing navigation from one state to another, keeping the saved settings or setting the default values in case of absence. On the left, however, there is the navigation bar of the single macro-topic (in this case the navigation bar of the "Analysis" status is shown), with all the possible actions for that status following the order from top to low for a sequential exploration. It also allows the UNDO operation, to allow the user to go back to the previous step and reformulate a new hypothesis.

In reality, the user can also pass from one activity to another as the tool allows saving the status of the single page.
Finally, the central structure is organized always showing the title of the current page which indicates to the user his position within the analysis, and in the central part of the page we have the elements indicative of the specific activities that can be performed. For example in the case of the Figure 5.25 we find a table of the top-K with selectable rows, a selection bar to indicate how many rows to display, a scroll bar to indicate the range of rows to be displayed and finally a stepper, i.e. an indication numeric of the steps to be performed (selection of the row of interest, Shapley Value graphic display of the selected row, latex graphic display of the selected row).
The user is never treated as an oracle, he will not be able to change relevant details of the method used but at the same time allowing a series of interactive actions we can involve him in the analysis process, enhancing those aspects that we consider interesting in the evaluation, with a dynamic process and results fast, meaningful and above all online (just a change to a parameter that the page updates automatically accordingly).

5.3 Description of interactive operations

Described the data exploration techniques for classification analysis (Chapter 4) and analyzed the choices relating to the design structure (5.2.3), we focus in this section on the relationships between the theoretical part and the graphics. We will analyze which operations are available to the user for each important step of the analysis, starting directly from the beginning of the analysis without including the initial pages already described (cit figure). We can divide our analysis into four basic macro steps: 1) Choice of the dataset, 2) Selection of metrics, 3) Analysis, 4) Global evaluation.

5.3.1 Choice of the dataset

The first macro step concerns the choice of the dataset presented to the user with a view containing two sections. The first section (D) allows the user to select one of the proposed datasets in the proposed table (A) and to view (C) a preview of the selected dataset (the first 5 instances will be shown), furthermore, the name will be displayed at the top of the title of the selected dataset (B).

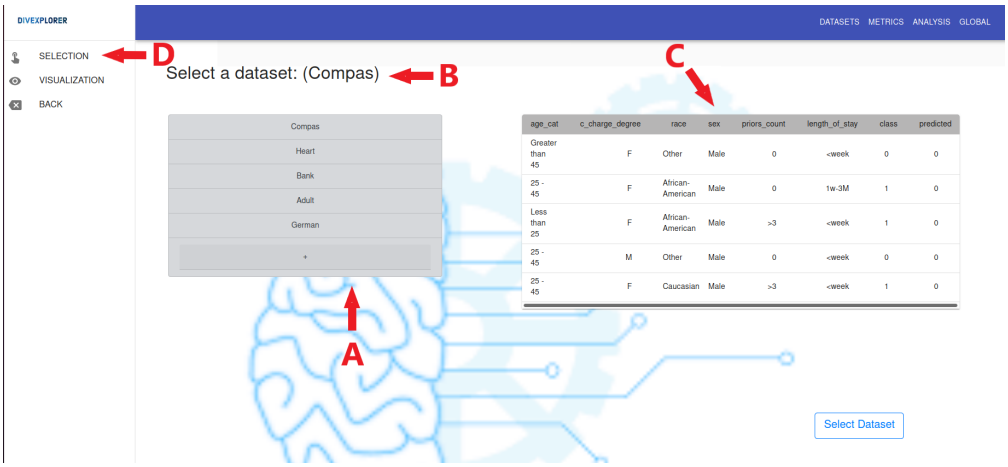


Figure 5.26: View of Dataset Choice

If the user clicks on the box relating to another dataset with respect to the default value, the view will be updated automatically, signaling to the user the change of state within the application through a backdrop that indicates the dataset value next to the name and, once loaded, showing its preview (Figure 5.27).

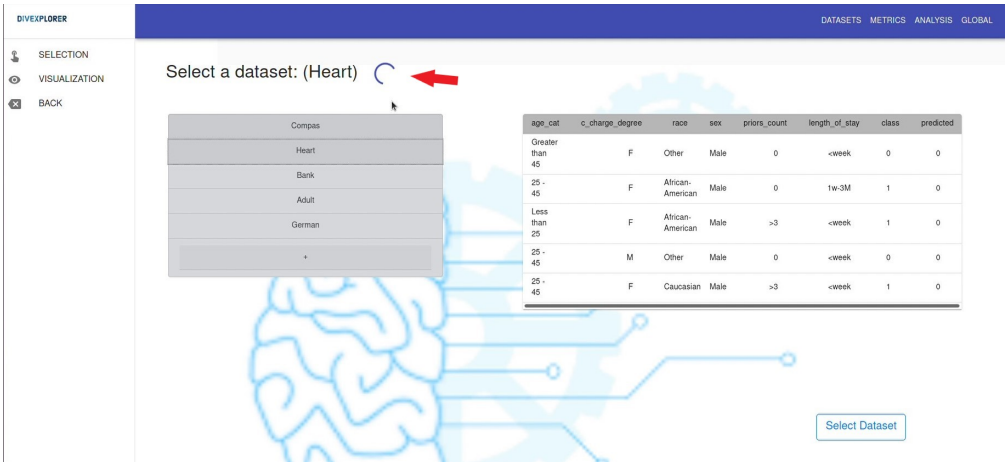


Figure 5.27: Loading the new dataset into the application

By clicking on the button with the "+" symbol in the list of datasets, the user can load a dataset of his interest. A new box will open where the user can select the dataset by pressing the button with the magnifying glass (A) which will open the display of his file system to facilitate the operation. Once selected it will display the name in the text box (C) and pressing the "Select" key (B) will confirm the loading of the dataset within the application.

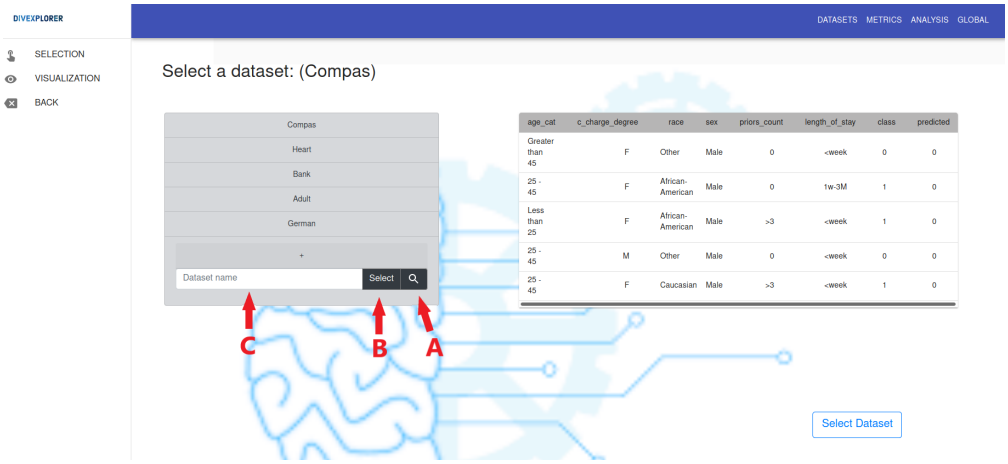


Figure 5.28: Insertion of a dataset chosen by the user

To pass to the second section, the user just has to use the navigation bar and click "Visualization". In the second section, the user can explore the dataset selected in the first section completely. The name and the table containing all the dataset partitioned into multiple pages are shown that the user can explore using the commands at the end of the table (the user can select the number of rows to be displayed on a page and can scroll through the various pages by simply clicking on the arrows).

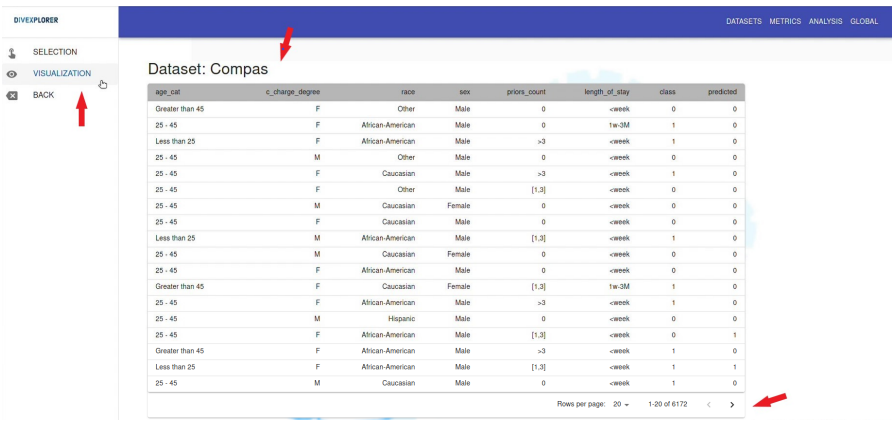


Figure 5.29: Display of the entire dataset selected by the user

The confirmation, by the user, of the choice of the dataset by clicking on the "Select Dataset" button brings up a dialog for the set of analysis setting parameters. The user can select, through two drop-down menus, the true class and predicted class, or leave the default values, and as we have analyzed in Chapter 4, he must select the value of the support threshold with which we will identify the frequent subgroups through the Frequent Pattern Mining algorithm. The user can also decide to enter only the true class, in which case he will not have all the proposed metrics available in the next phase as it will not be possible to compare the two classes.

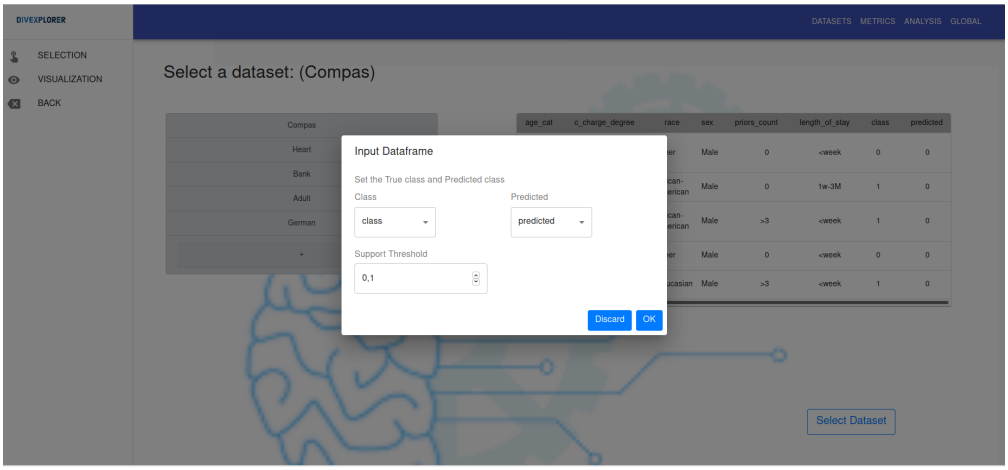


Figure 5.30: Setting the parameters of the analysis

At this point, the user can decide to go back and make another choice by pressing the "Discard" button, or confirm the settings by pressing the "OK" button and move on to the next step.

5.3.2 Selection of metrics

The application will save all the states related to the choice of the selected dataset and will show the user the view for the second macro step: the selection of the evaluation statistics.

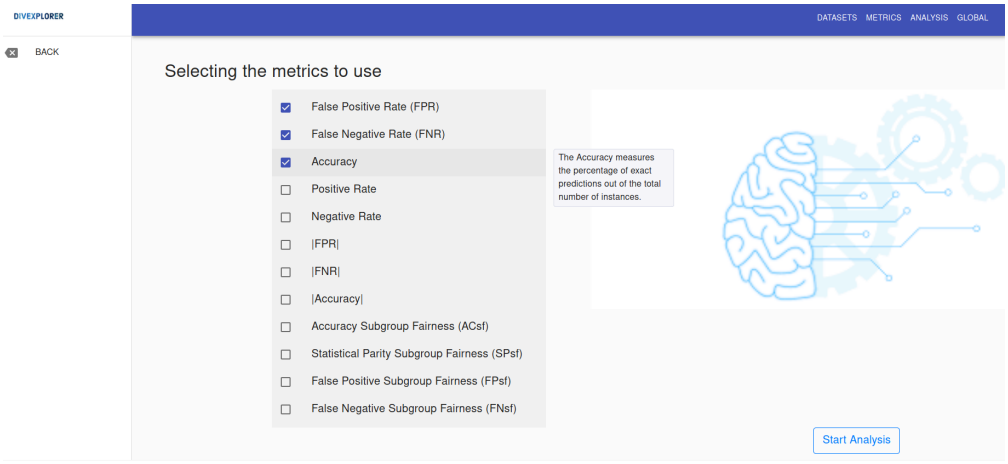


Figure 5.31: Selection of evaluation statistics

The selection of metrics is presented through a list of proposals each with its checkbox. Multiple selections are foreseen, the user can select all the metrics of his interest. To help the user in the selection of the metrics that he considers of particular interest, by hovering the mouse over the name of a metric, a tooltip is displayed on the side, that is a small suggestion containing the summary description of the metric. In this way, the user does not make a random choice but can quickly understand what that statistic is analyzing and decide whether to check or not. Once all the metrics of interest have been selected, the user clicks the "Start Analysis" button, the application saves the status relating to the metrics and directs the user to the next stage.

5.3.3 Performing the Analysis

Evaluating Frequent Patterns

The analysis begins with the evaluation of frequent patterns presented to the user in the form of a table. Above the table, there are as many Tabs as the metrics selected in the previous step. In this way, the user can easily navigate between one metric and another to evaluate the corresponding values. The table is constructed showing the support value, the composition of the item, the confusion matrix, the length of the item, the count of instances that satisfy that group, and the values relating to the metric selected in the Tab.

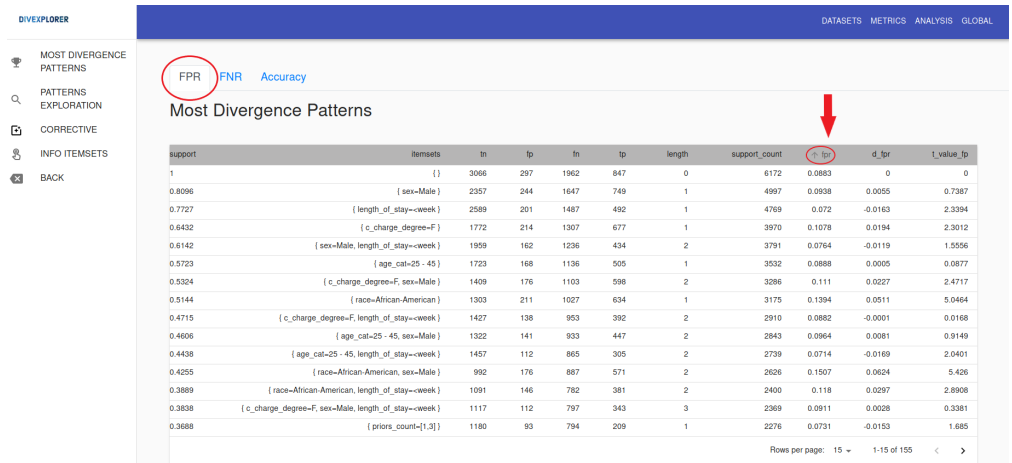


Figure 5.32: Evaluation of the most discrepant patterns

The metric selected in the Tab is the one taken into consideration in all the subsequent phases of the analysis, to change it, the user must return to the view relating to frequent patterns and change the selected Tab. The table allows a series of manipulations, for example by placing the mouse on a column value an

arrow will appear, by clicking on it it will be possible to sort the table according to the selected column value. Clicking a second time on the arrow will change the direction of the arrow (if it points up the order will be ascending, if it points down the order will be descending). The table contains all the calculated instances, it is possible to change the number of rows to be displayed using the appropriate command at the end of the table. The user can navigate through the other pages of the table by clicking on the arrow at the bottom, the count of the pages being viewed is also shown compared to the total.

Detailed exploration of the patterns

Following the navigation bar relating to the operations possible for the macro topic (the bar on the left), the next step allows the exploration of the patterns in detail. To guide the user during this exploration we propose (Figure 5.33) the display of a Stepper (C). The Stepper is used to indicate an indicated procedure, showing the numbered steps, showing each time the completed steps. The first step involves selecting a row in table (A) which indicates the corresponding group discrepancy value for each itemset. The table shows the first K elements with the largest group discrepancy, the user can decide to change the value of K to view more rows simply by changing the value in the numerical box (B).

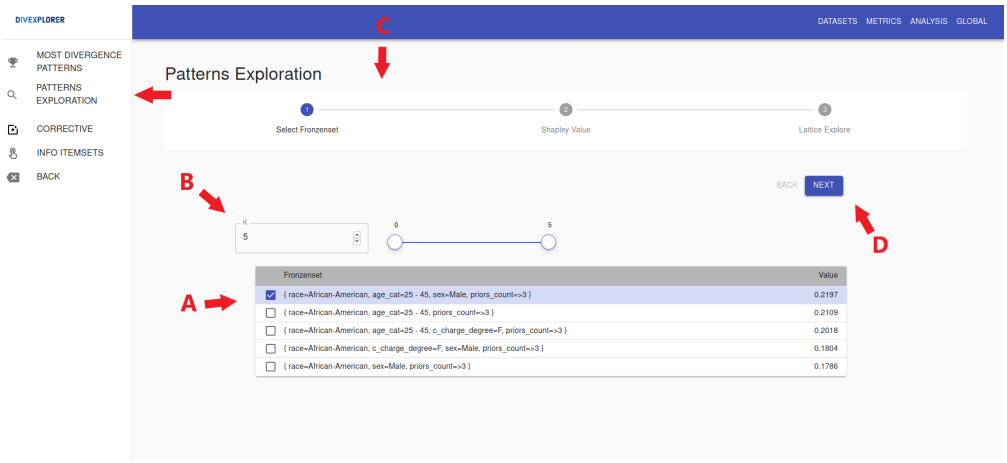


Figure 5.33: Exploration of most discrepant patterns: pattern selection

If the user selects too high a number of K, the table grows so that it can no longer be represented on a single page and it is necessary to scroll down to view the final values.

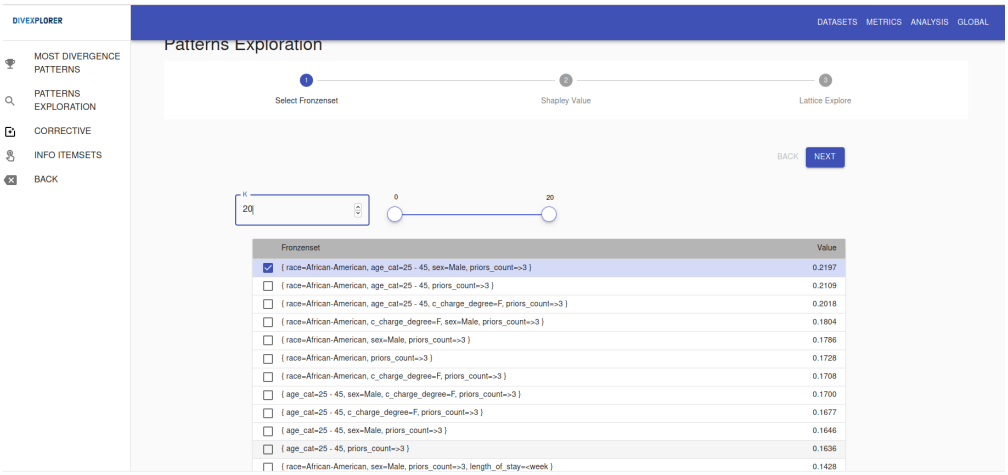


Figure 5.34: Exploration of most discrepant patterns: top 20 pattern

To avoid having to go up and down to follow the steps we have inserted a Slider that indicates the range of itemsets to be displayed. In the case of 20 rows, for example, we can decide to display from the fourth to the fifteenth row, obtaining a table perfectly suited to the page.

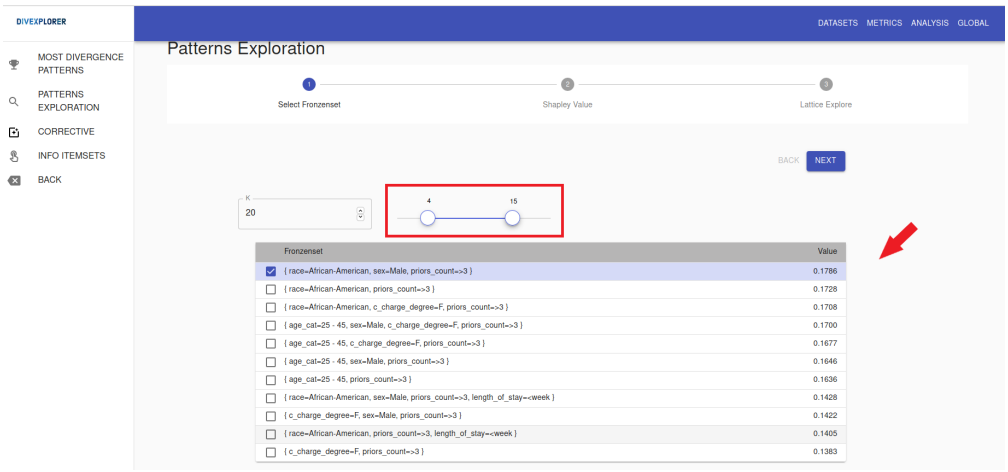


Figure 5.35: Exploration of most discrepant patterns: select a range of patterns to display

The user selects an itemset to evaluate it in detail, we consider the first available value, by pressing the “Next” button we move on to the next step of the exploration. The Stepper indicates that we are in the second step, by coloring the step with the number 2.

In the second step, we have the evaluation of the local contributions of the individual attributes that make up the item set.

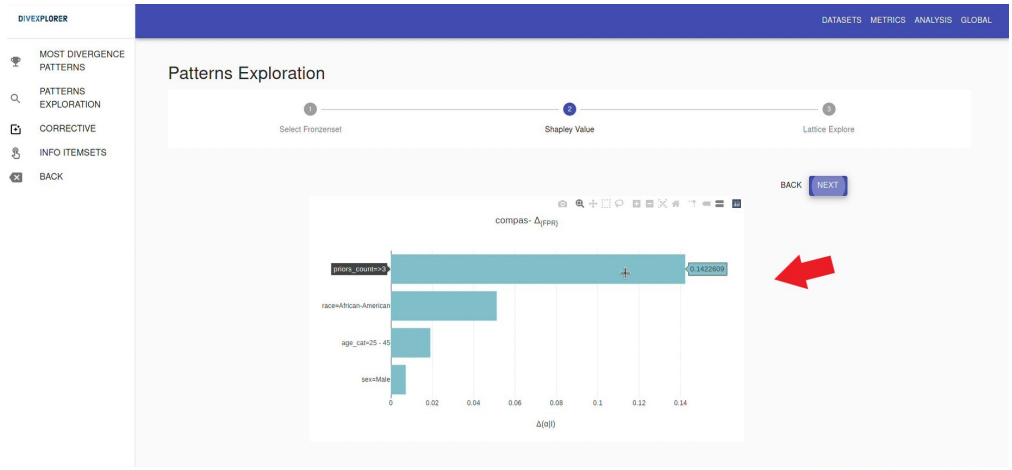


Figure 5.36: Exploration of most discrepant patterns: Shapley Value of selected pattern

The Shapley Value is represented as a horizontal bar graph. The graph allows the display of the precise value by simply passing the mouse over the bar of interest. This facilitates the evaluation of the results, not forcing the user to search for the correct value by enlarging the image and estimating the result based on the values of the grid. Simply position the mouse and the corresponding value will be immediately displayed in the form of a label on the side of the bar.
By pressing the "Next" button again we complete the second step and move on to the last step. The third step allows the interactive exploration of the pattern through a visual representation of the network of the group of objects. The nodes of the grid represent a subset of the selected item starting from the empty subset and adding an item to each level up to the last level that represents the item itself. The group discrepancy value is indicated for each subset.

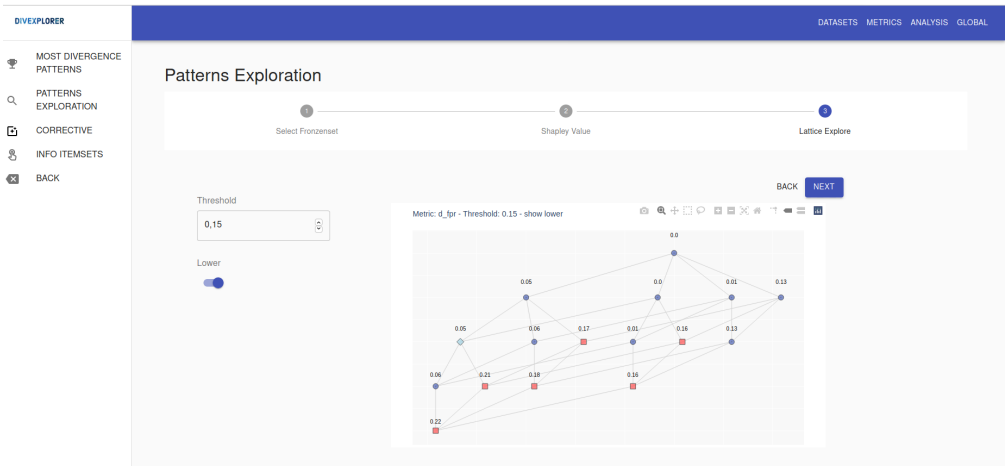


Figure 5.37: Exploration of most discrepant patterns: Lattice Search

The user can select the threshold value and observe which subsets have a higher (red squares) or lower (blue circles) value. With the lattice search, it is also possible to verify which items exhibit a regulatory behavior, that is, in the lattice descent, they tend to lower the group discrepancy value. The items with the regulatory behavior are displayed using blue diamonds. The user can decide whether to view or not view these items simply by activating or deactivating the Switch indicated with "Lower".

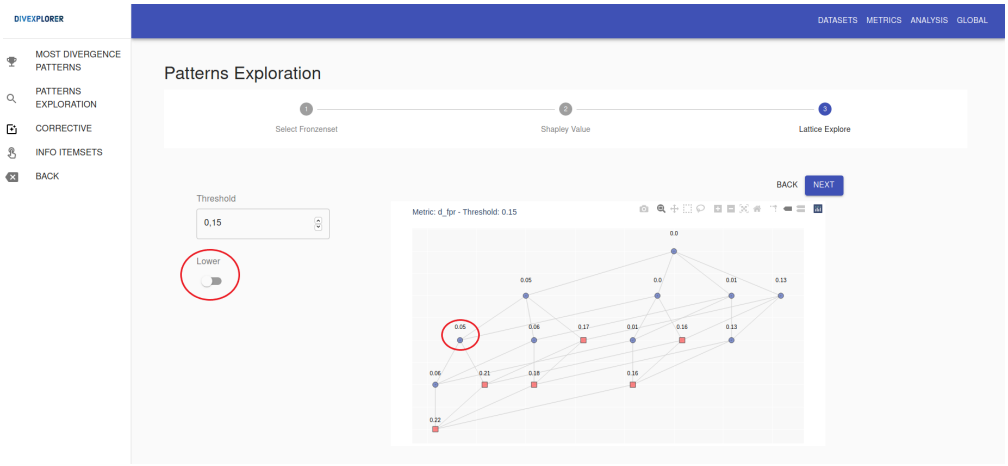


Figure 5.38: Exploration of most discrepant patterns: Lattice Search without displaying the regulatory items

Search for regulatory items

The search for regulatory items is presented to the user through a table that collects an itemset for each row, the item that is added, the value of the item set statistic, and the variations that the added item causes to the itemset. The user can select how many lines to display (the default value is 5) and can select, as he wishes, one line at a time to graphically check the influence that the added item has on the item set.

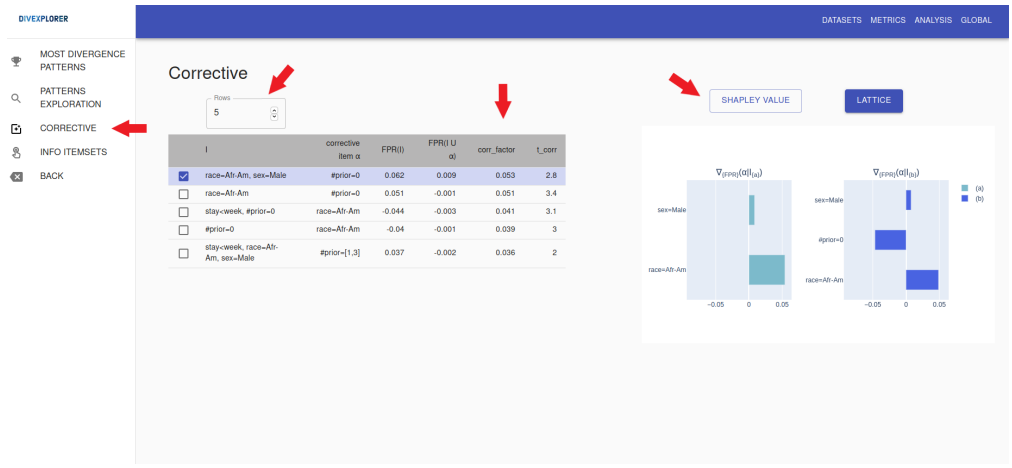


Figure 5.39: Search for items with a regulatory effect

We propose to the user the visualization of the comparison chart between the Shapley Value of the original item and that of the item set with the regulating item, and the representation of the lattice search of the adjusted item. To carry out these operations two buttons are indicating the name of the graph to be displayed. The button with a light background indicates that the user has clicked on it. If the user decides to view the representation of the lattice search by clicking on the "Lattice" button, the necessary settings will appear to change the threshold and to view the regulating items as in the case of Figure 5.41.

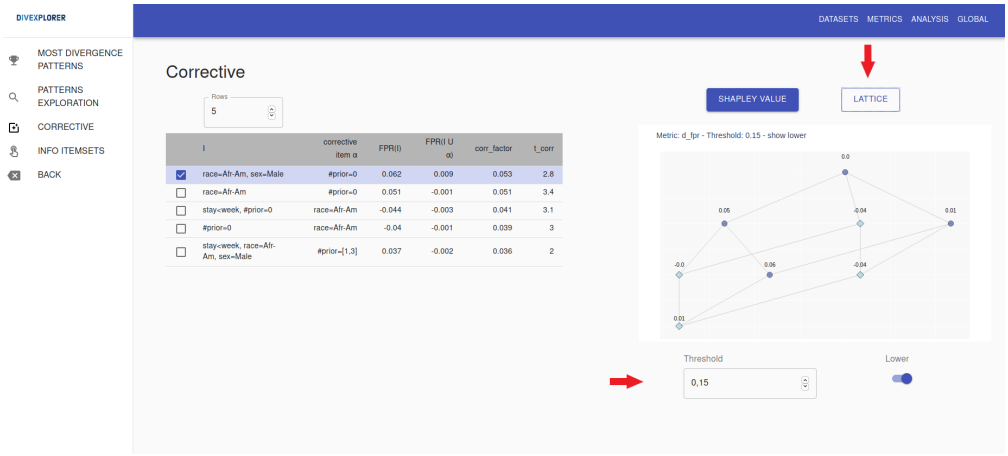


Figure 5.40: Lattice Search for items with a regulatory effect

Displays information patterns

The last phase of this analysis allows the user to trace information relating to a pattern of interest. The user has a list of Selects available, one for each attribute of an instance.

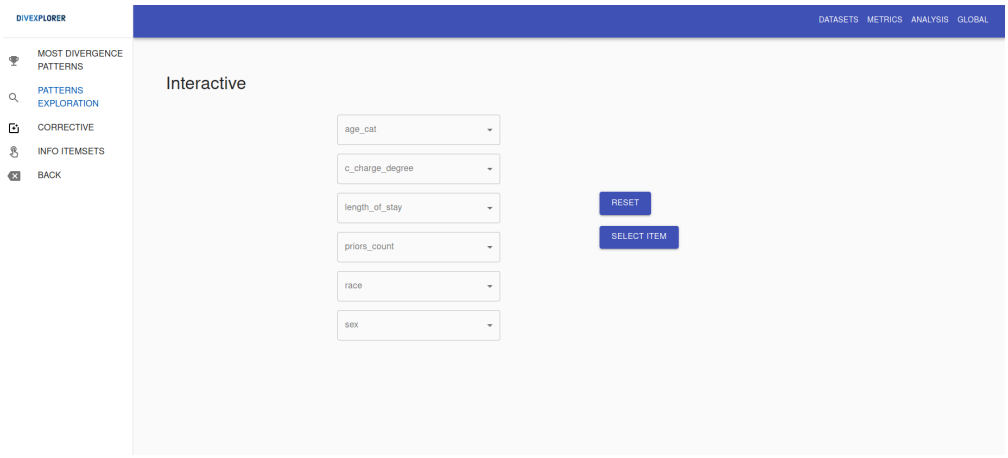


Figure 5.41: Search for information about an itemset by selecting attribute values

The user can select a value for each attribute, not necessarily there must be a value for each attribute, and clicking on the "Select Item" button will display the information relating to the selected subgroup. A row containing the support value, the item set composition, the confusion matrix, the statistic value, and the group discrepancy value will be shown.

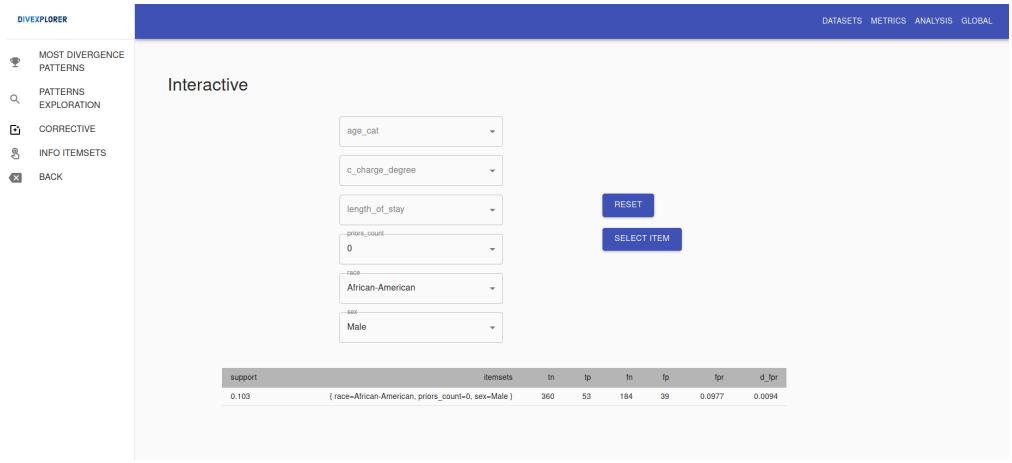


Figure 5.42: Search for information about an itemset with itemset found

The user can formulate different combinations to obtain the information. The “Reset” button is used to reset the Select values, avoiding having to manually set each unsolicited attribute to the initial default value of “None”. If the user selects an itemset that has a support value lower than the threshold specified at the start of the analysis, a Dialog appears that alerts the user of the situation and asks if the user still wants to view the information, in which case it will confirm with the "Agree" button, or if he prefers to go back and request a new combination (by pressing the "Disagree" button).

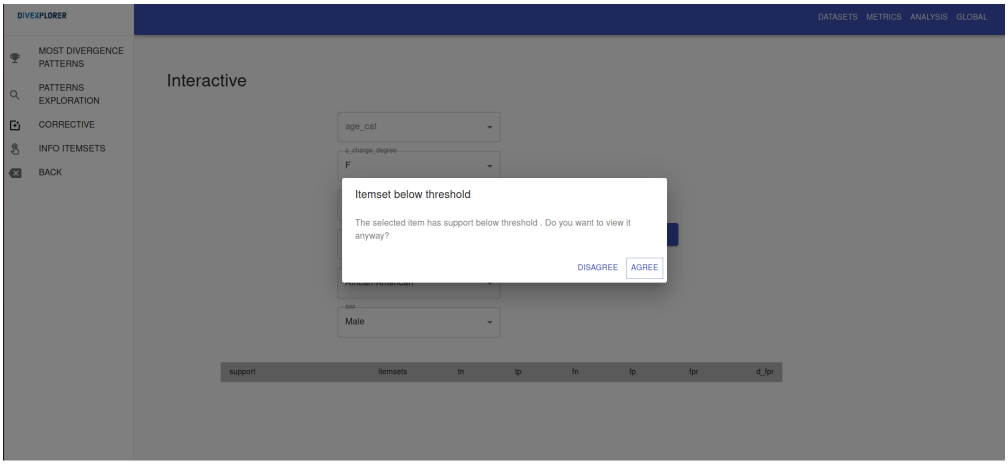


Figure 5.43: Search for information about an itemset with itemset not found

Finally, with the "Back" button on the navigation bar, it is possible to return to the previous phase of the analysis or, the selection of metrics.

5.3.4 Global Evaluation

The last macro step is the evaluation of the global results which the user can access through the main navigation bar by clicking on "Global". The overall results contain a summary of the entire analysis and can help the user to better identify the correlations between items and how these affect the group discrepancy value. The home page shows the user the various global rating options enclosed in 4 clickable sections diversified by color.

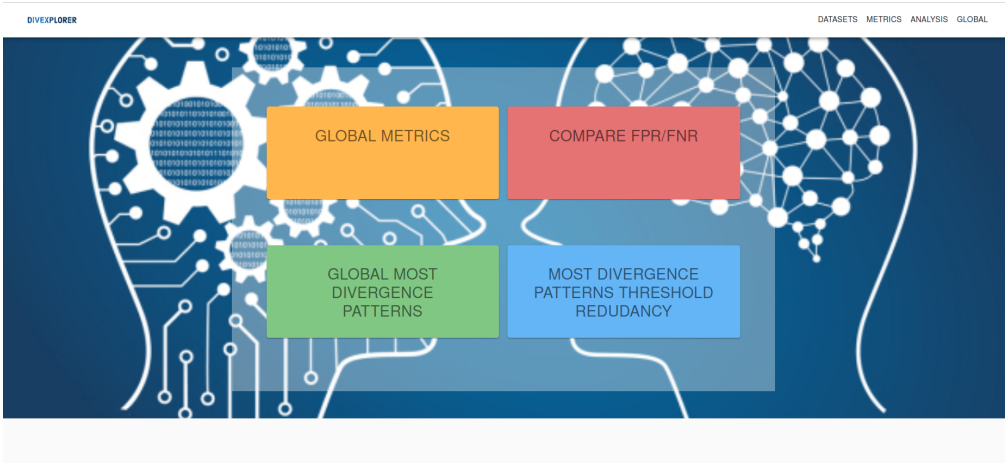


Figure 5.44: Display of options for global evaluations

Global Metrics

The first option allows the evaluation of global results by single metric (the green section is very similar, it presents the global contributions of the first K elements, with K equal to Figure 5.33). The user can select the metric of interest through the Tabs at the top of the page and evaluate the global Shapley values for each item.

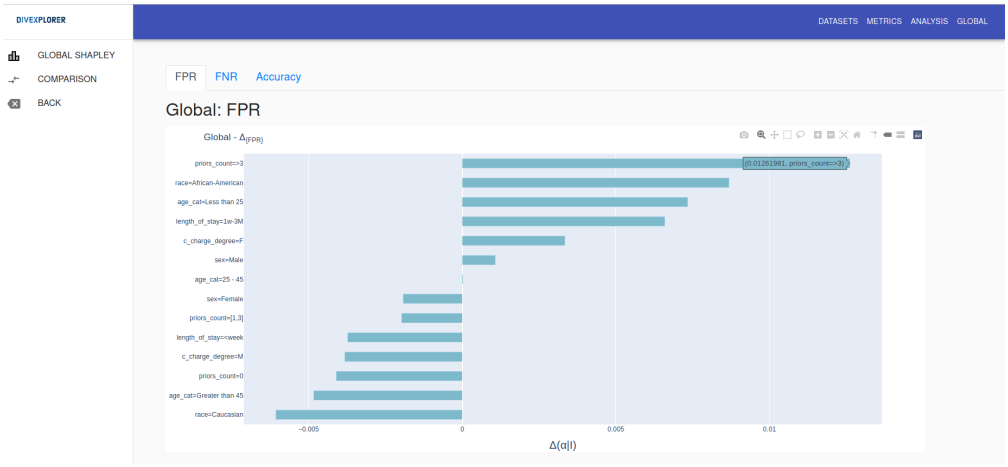


Figure 5.45: Evaluation of the global Shapley values for each item

The user can also simultaneously evaluate the global Shapley values and the respective group discrepancy values of the selected metric by navigating in the sidebar to the second option. The visualization presented as in the Figure 5.46 allows the user to verify which coalitions of items lead to a high group discrepancy value.

Figure 5.46: Evaluation of the global Shapley values and group discrepancy values for each item

Compare FPR and FNR

In this section, the user can compare the global Shapley values relating to the False Positive Rate and the False Negative Rate, presented as two bar graphs with different colors placed side by side. The user can evaluate the different global contributions of the items for the two statistics. Furthermore, the parallel representation allows a greater understanding of the correlations of items that influence the final value tending to raise or lower it.

Figure 5.47: Evaluation of the global Shapley values for False Positive Rate and False Negative Rate

Most discrepancy patterns

The fourth section (Blue color Figure 5.44) allows the user to view a summary of the K models with the highest group discrepancy value. The display is presented to the user in the form of a table containing for each instance the value of the support, the composition of the item set, and the values relating to the metric taken into consideration.

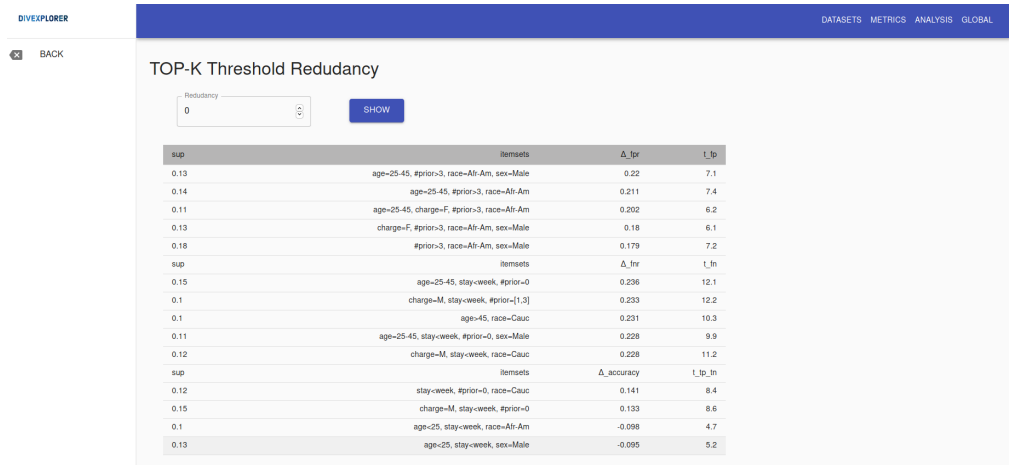


Figure 5.48: Evaluation of the most discrepant patterns with threshold for discarding itemset with irrelevant variation

As you can see from the Figure 5.48 there is a text box in which to enter a threshold value defined as the redundancy threshold. The redundancy threshold is used to discard itemsets whose group discrepancy variation with the addition of an item is not particularly relevant (less than or equal to the indicated threshold). This view helps the user to better identify in which subgroups there may be some inefficient performance in terms of classification. The user can change the redundancy threshold simply by changing the number in the text box, to avoid a continuous update for each digit entered, the result with the changed threshold will be displayed after clicking the "Show" button.

User Experience Considerations

Considering the user-experience, the interactive tool presented is extremely easy to use. Its strength lies in many factors such as:

1. the organizational structure of the components, very common and easy to understand;
2. quick update of the result based on parameter changes;
3. division of the analysis states in a decisive way, which does not confuse the user;
4. saving intermediate results, allowing you to go back without losing information;
5. it does not require closing and reopening to restart the analysis from scratch, just return with the navigation bar to the choice of the Dataset;

6. allows the undo of operations without having to be updated;

User involvement is the main aspect that we have tried to achieve in every single operation, even where the analysis did not include modifiable parameters we have tried to keep the user's attention active by allowing, for example, the evaluation of interactive graphs. On each page the user can act, making the process active.
The thesis aims to convey the idea that an approach implemented in this way can be of greater help in the wide search for optimization of machine learning systems, also importing a more aesthetic aspect into the study that could simplify the development and understanding.

Chapter 6

Experiments

In this Chapter, we will show experiments of the results that can be obtained with the interactive tool presented in Chapter 5. In the first section the experiments carried out using the tool will be presented, in the second section, a comparison with another will be presented interactive system, FairVis[18].

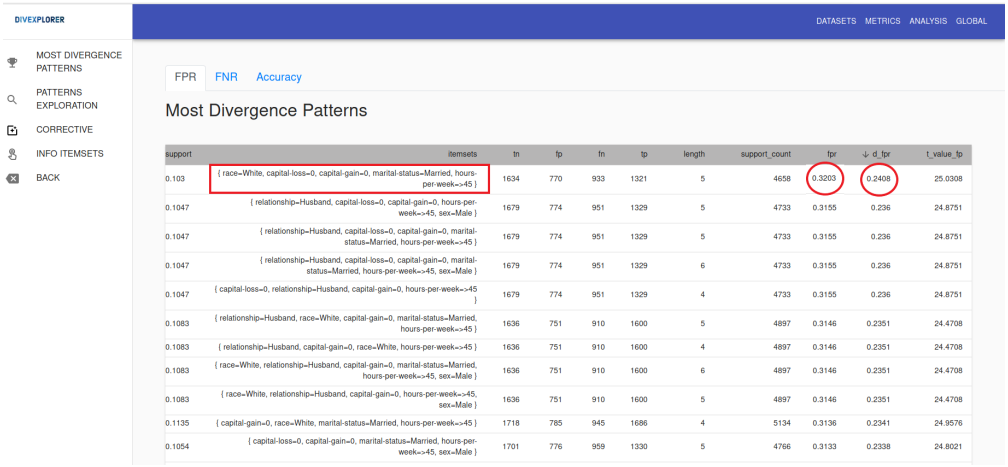
6.1 Evaluation of the results of the interactive tool

We analyze the Adult Dataset to determine if a person earns more than 50K per year. We divide the analysis into three sections by considering three statistics, False Positive Rate, False Negative Rate, and Accuracy.

6.1.1 False Positive Rate (FPR)

Analysis of the most discrepant patterns and the contributions of attributes

Let's consider the pattern analysis with a higher group discrepancy value than the overall (for False Positive Rate equal to 0.0795). We search for the pattern that has the greatest group discrepancy value, i.e. that deviates most from the general behavior and we find, by ordering by decreasing group discrepancy values, that the pattern that has the greatest distance is the one composed of race = White, capital-loss=0, capital-gain = 0, marital-status = Married, hours-per-week=>45.



support	itemsets	tn	fp	fn	tp	length	support_count	fpr	↓ d_fpr	t_value_fp
0.103	{ race=White, capital-loss=0, capital-gain=0, marital-status=Married, hours-per-week=>45 }	1634	770	933	1321	5	4658	0.3203	0.2408	25.0308
0.1047	{ relationship=Husband, capital-loss=0, capital-gain=0, hours-per-week=>45, sex=Male }	1679	774	951	1329	5	4733	0.3155	0.236	24.8751
0.1047	{ relationship=Husband, capital-loss=0, capital-gain=0, marital-status=Married, hours-per-week=>45 }	1679	774	951	1329	5	4733	0.3155	0.236	24.8751
0.1047	{ relationship=Husband, capital-loss=0, capital-gain=0, marital-status=Married, hours-per-week=>45, sex=Male }	1679	774	951	1329	6	4733	0.3155	0.236	24.8751
0.1047	{ capital-loss=0, relationship=Husband, capital-gain=0, hours-per-week=>45 }	1679	774	951	1329	4	4733	0.3155	0.236	24.8751
0.1083	{ relationship=Husband, race=White, capital-gain=0, marital-status=Married, hours-per-week=>45 }	1636	751	910	1600	5	4897	0.3146	0.2351	24.4708
0.1083	{ relationship=Husband, capital-gain=0, race=White, hours-per-week=>45 }	1636	751	910	1600	4	4897	0.3146	0.2351	24.4708
0.1083	{ race=White, relationship=Husband, capital-gain=0, marital-status=Married, hours-per-week=>45, sex=Male }	1636	751	910	1600	6	4897	0.3146	0.2351	24.4708
0.1083	{ race=White, relationship=Husband, capital-gain=0, hours-per-week=>45, sex=Male }	1636	751	910	1600	5	4897	0.3146	0.2351	24.4708
0.1135	{ capital-gain=0, race=White, marital-status=Married, hours-per-week=>45 }	1718	785	945	1686	4	5134	0.3136	0.2341	24.9576
0.1054	{ capital-loss=0, capital-gain=0, marital-status=Married, hours-per-week=>45, sex=Male }	1701	776	959	1330	5	4766	0.3133	0.2338	24.8021

Figure 6.1: Most discrepant patterns for Adult Dataset (FPR)

Our interactive tool allows us to go into the detail of the group discrepancy value, for this reason, we select the row corresponding to the pattern with the greatest discrepancy value and analyze the local contributions of the attributes that make up the pattern.

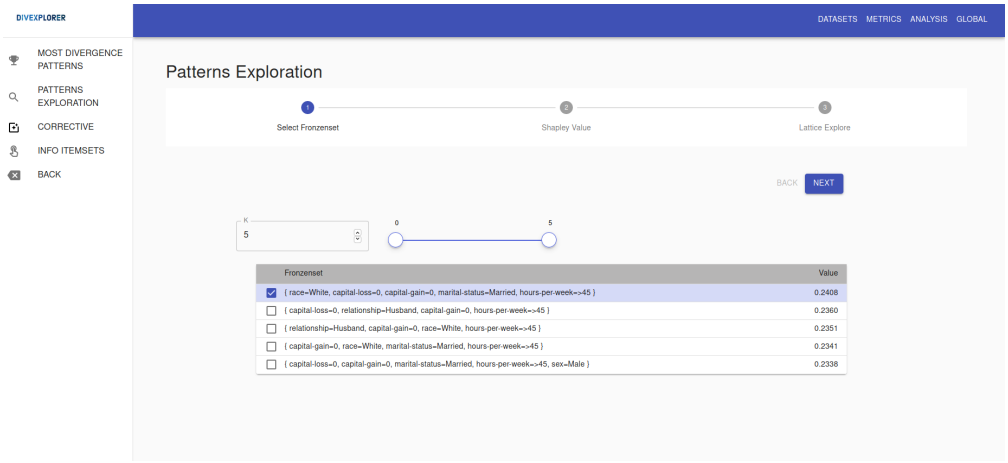


Figure 6.2: Selection of most discrepant pattern for Adult Dataset

The Shapley values for the pattern under consideration show that the attribute values that contribute most to a high group discrepancy value are married marital status, followed by hours-per-week = >45. The presence of these attributes makes the prediction tend towards a false positive.

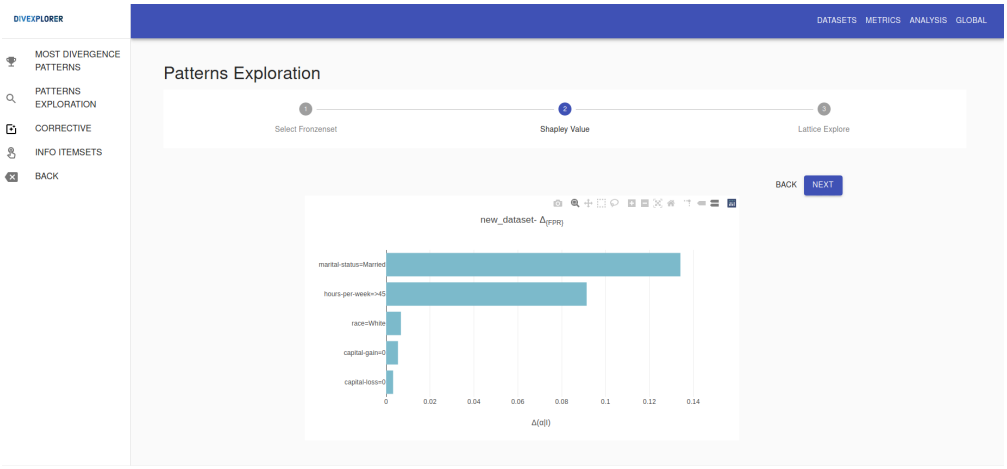


Figure 6.3: Shapley values for most discrepant pattern for FPR (Adult Dataset)

This concept is made even better by the visualization of the lattice search, in fact, at the third level, we already find the first subgroup that exceeds the threshold and is composed only of the two attributes that stand out most in the Shapley values. The addition of the other attributes does nothing but increase the group discrepancy value, reaching the last level which represents the initial pattern with the maximum value.

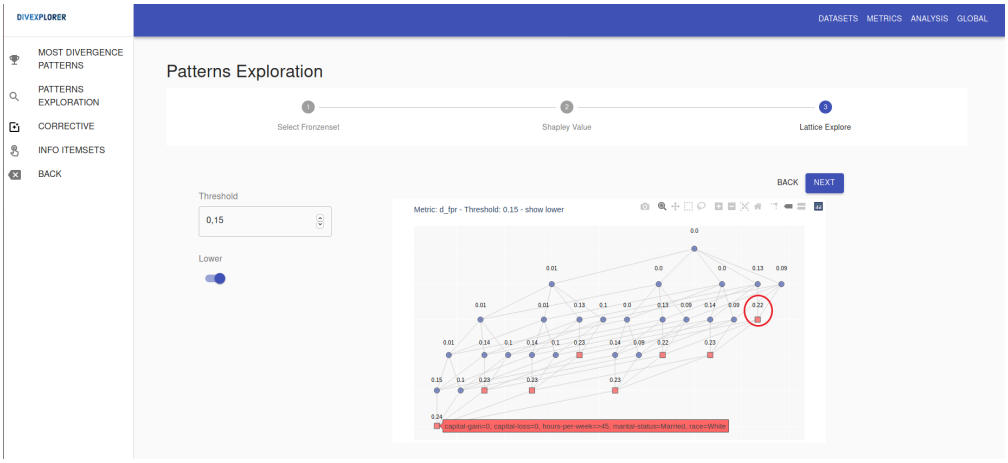


Figure 6.4: Lattice Search for most discrepant pattern for FPR (Adult Dataset)

Search for regulatory items

To search for the regulating item we consider the first five instances proposed by the tool, from the tabulated values we note that the addition of the item occupation = Blue-Collar has a damping effect on the group discrepancy value, balancing the strong contribution that the relationship attribute equal to Husband gives in the opposite direction tending to raise the group discrepancy value. The concept is best expressed by evaluating the Shapley values of the situation before and after adding the item.

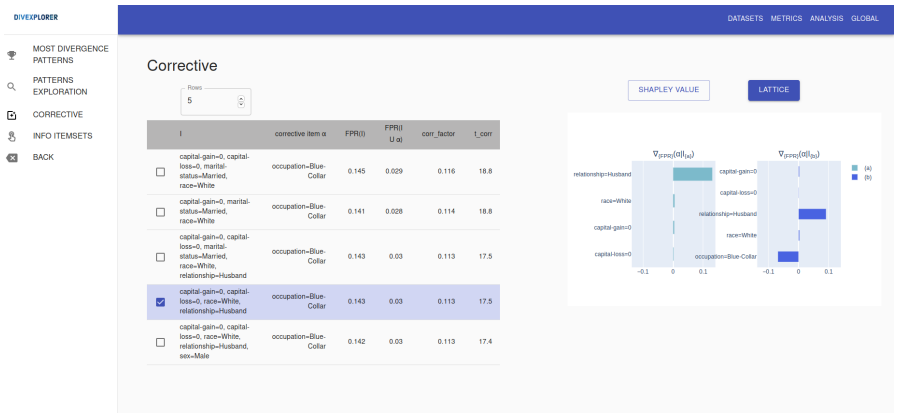


Figure 6.5: Regulatory item search for FPR (Adult Dataset)

The lattice search emphasizes the concept of regulation even more, in fact for the case considered we can note that the addition of the item occupation = Blue-Collar leads to the creation of a pattern whose subgroups all have group discrepancy values lower than the threshold set and an entire half of the lattice forms subgroups with discrepancy values that decrease as you go down the lattice.

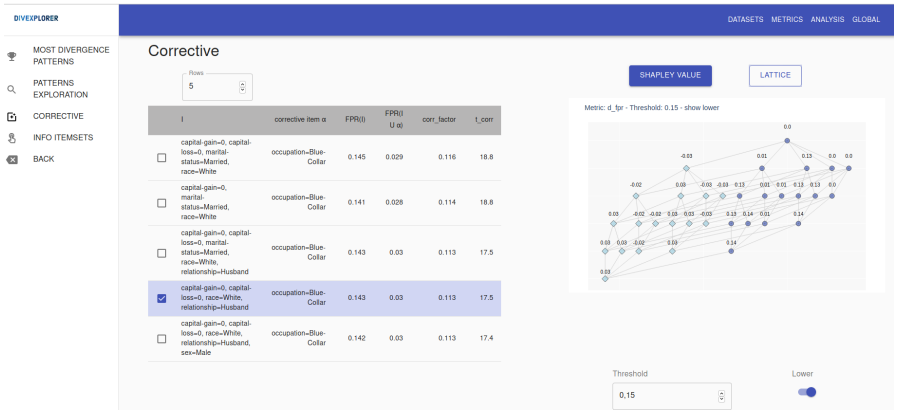


Figure 6.6: Lattice Search of regulatory item for FPR (Adult Dataset)

Global Evaluation

From the evaluation of the global results we can determine that the attribute values that increase the group discrepancy value if related are marital-status = Married, relationship = Husband and hours-per-week = >45.

Figure 6.7: Global Discrepancy Group for every item for FPR (Adult Dataset)

6.1.2 False Negative Rate (FNR)

Analysis of the most discrepant patterns and the contributions of attributes

Let's analyze the model considering the False Negative Rate statistic with an overall equal to 0.3901. The evaluation of the most discrepant patterns ordered according to the group discrepancy values shows that there are different subgroups with the same maximum group discrepancy value (0.6099), we choose the itemset identified by hours-per-week = <40, education = High School grad, capital-loss = 0, capital-gain = 0, marital-status = Never Married.
This instance will be the one we will consider in the search for the local contributions of its constituent attributes.

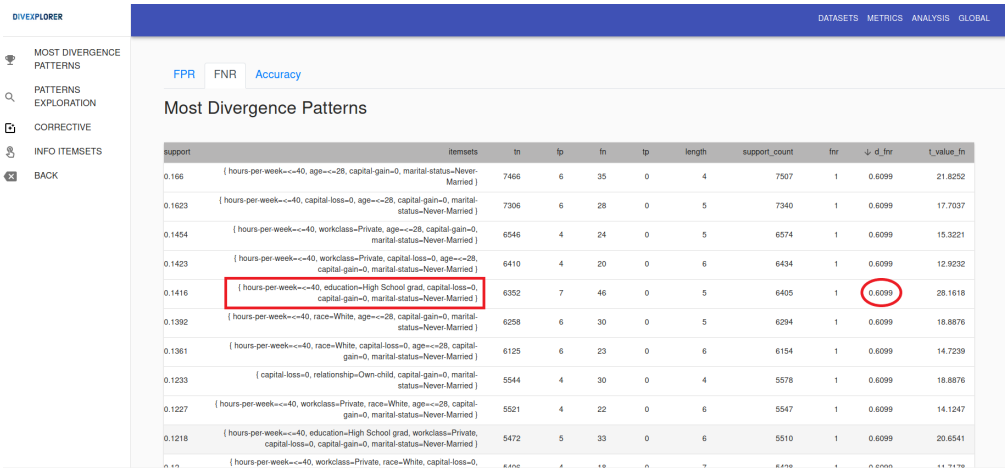


Figure 6.8: Most discrepant patterns for Adult Dataset (FNR)

The Shapley values clearly show that for the pattern under consideration, the attributes that contributed most to the high value of group discrepancy are in order the marital-status equal to "Never Married", the capital-gain equal to "0" and the education equal to "High School grad".

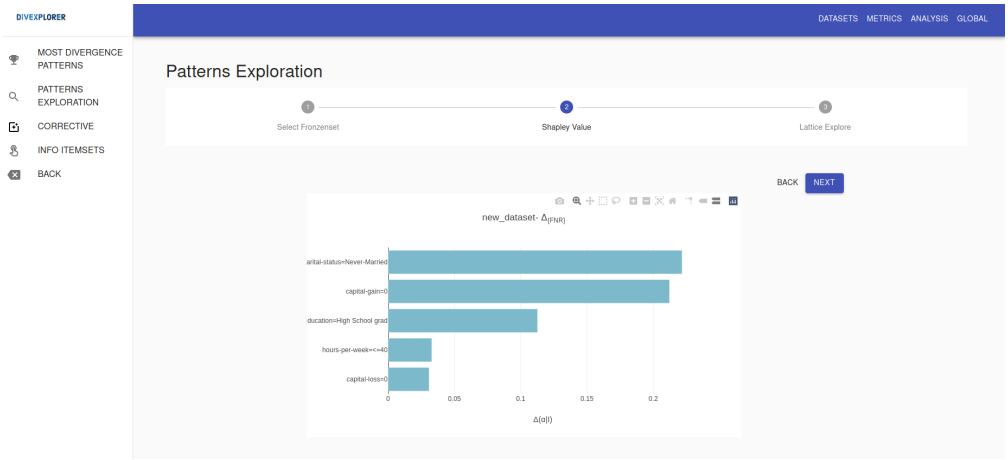


Figure 6.9: Shapley values for most discrepant pattern for Adult Dataset (FNR)

The evaluation of the lattice research is interesting in which the criticality of the pattern stands out, since already the first level (each node represents an attribute) there are subgroups that exceed the threshold. Furthermore, we can also note the presence of regulatory items. In the third level the subgroup composed of marital-status = Never Married and education = High School grad has a group discrepancy value equal to 0.27, the addition of the hours-per-week = <40 item lowers the group discrepancy value to 0.25 .

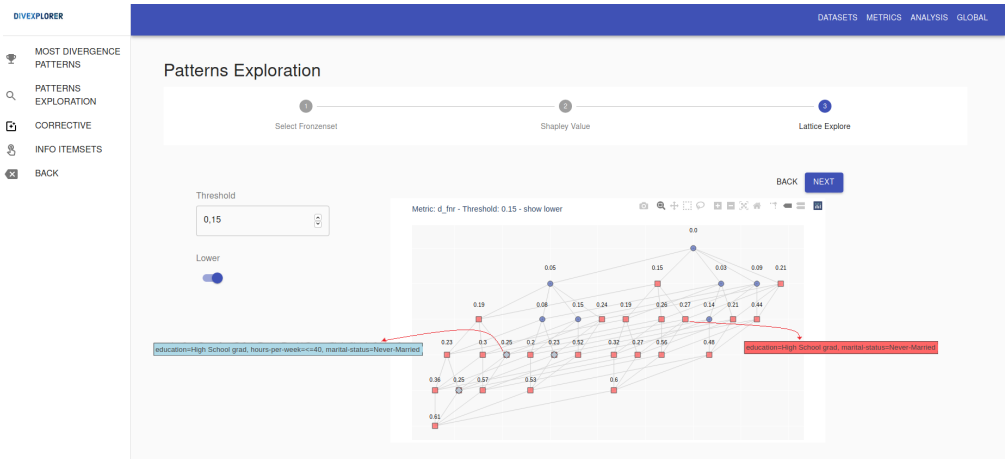


Figure 6.10: Lattice Search for most discrepant pattern for Adult Dataset (FNR)

Search for regulatory items

As we know, the search for items with a regulatory effect is more understandable in the view dedicated to this operation. In fact, the exploration can take place at a more detailed level because we present both the value of the statistic and its variation with the addition of the item and we also give a quantification of the regulatory effect. We have seen how in the pattern examined the attribute marital-status = Never-Married gave the greatest contribution to the high value of group discrepancy, we show the case in which this attribute instead has the opposite effect, it does not contribute by increasing the value of discrepancy but adjusts the effect of the others attributes by lowering the final value.
In the case of the False Negative Rate, we note that the attribute with the greatest contribution for the selected item is age <= 28 followed by education = High School grad. The addition of the item marital-status = Never Married adds an opposite contribution, even if not so significant compared to the contribution of age <= 28(the Shapley values highlight this situation), such as to lower the group discrepancy value (from 0.377 to 0.11) with a regulatory effect equal to 0.267.

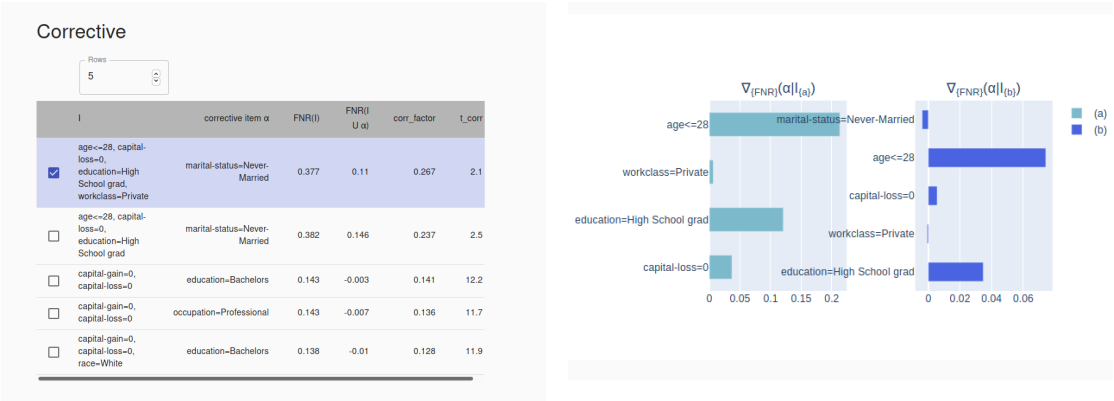


Figure 6.11: Regulatory item search for FNR (Adult Dataset)

The lattice search of the considered pattern better shows the regulatory effect due to the addition of the item marital-status = Never Married. Starting from a global perspective, we note the presence of many subgroups of the pattern that exceed the threshold of 0.15. The regulating effect already manifests itself at the third level while maintaining a high group discrepancy value (0.25). The evaluation for each level of the lattice shows how from the third level onwards the subgroups with a decreased value of group discrepancy gradually increase, obtaining also subgroups that return below the threshold value. The last level, for example, which represents the initial pattern, has a group discrepancy level equal to 0.11.

Figure 6.12: Lattice Search with the effect of the regulating item for Adult Dataset (FNR)

Global Evaluation

The global results provide a more general explanation of which attributes tend to raise the False Negative Rate value and indicate, thanks to the representation of the Shapley values, which coalitions of attributes contribute to dividing the behavior of the classifier from the general one. In this case we can conclude that the group of attributes that result in a high group discrepancy value for the False Negative Rate is composed of capital-gain = 0, education = Dropout , and marital-status = Never Married.

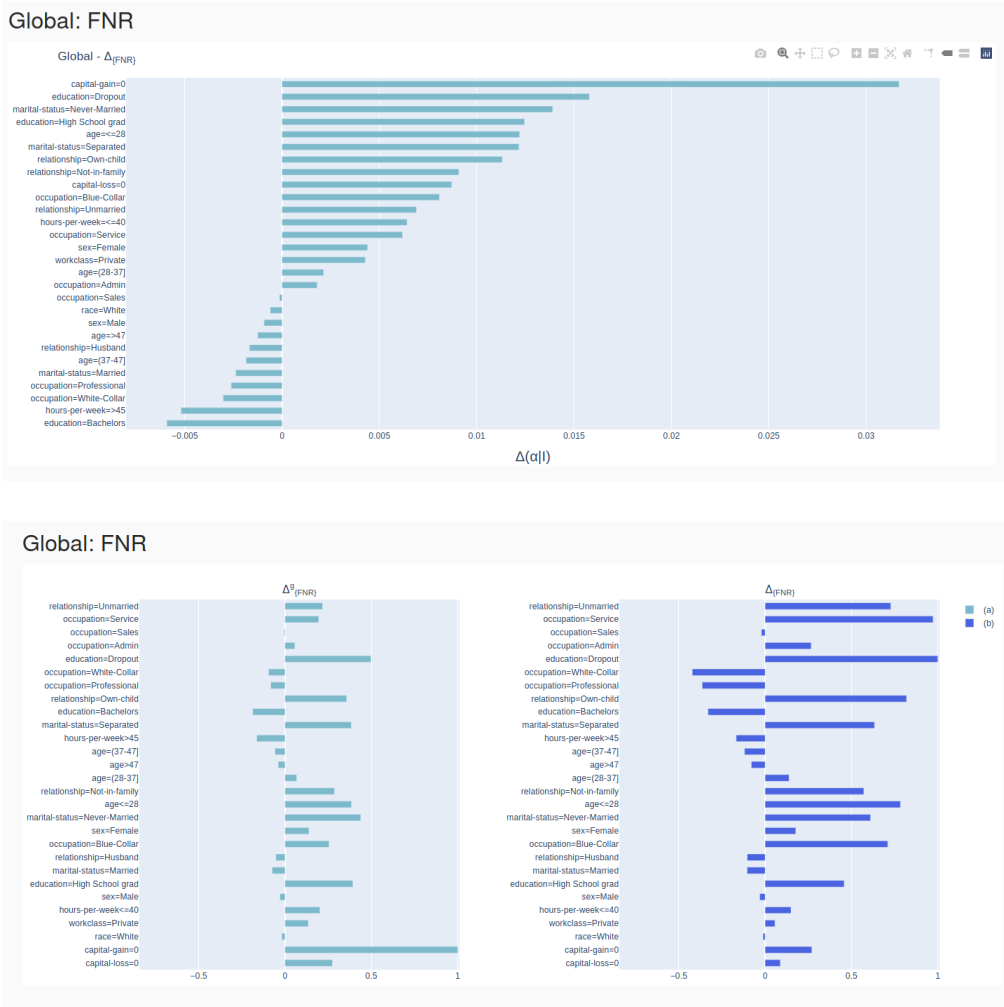


Figure 6.13: Global Discrepancy Group for every item for FNR (Adult Dataset)

6.1.3 Accuracy

Analysis of the most discrepant patterns and the contributions of attributes

Let us consider as a final example the evaluation of the model regarding the Accuracy statistic. We know that Accuracy measures the number of correct predictions on all predictions, so when evaluating the most discrepant patterns we will consider the subgroups with negative group discrepancy values (unlike the previous cases). A negative group discrepancy value indicates that in the subgroup the model is less accurate than the overall (group discrepancy value equal to 0.8435). We sort our table by descending values of group discrepancy and find that the itemset at greater distance (accuracy equal to 0.6344 and group discrepancy equal to -0.2091) from the overall is composed of race = "White", capital-loss = 0, capital-gain = 0, marital-status = "Married", and hours-per-week = >45.

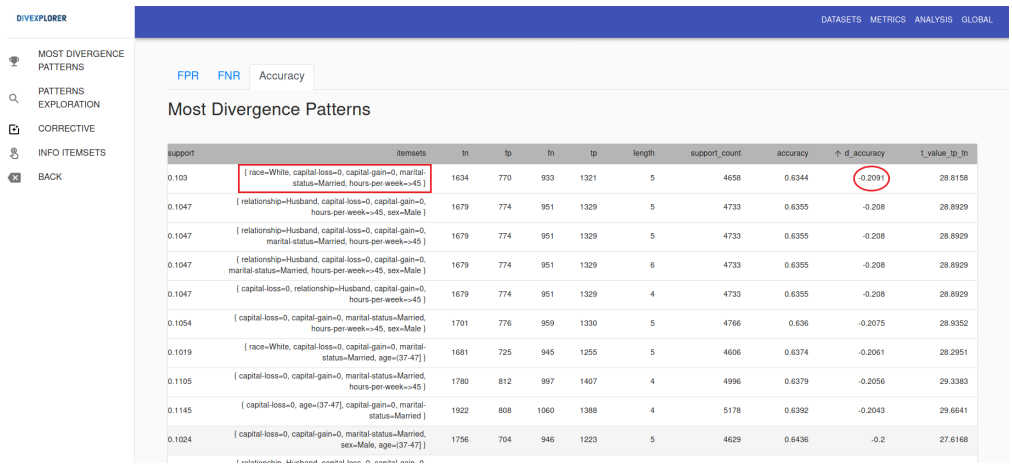


Figure 6.14: Most discrepant patterns for Adult Dataset (Accuracy)

Having obtained the pattern with the maximum group discrepancy value, we are interested in knowing for each attribute what is its individual contribution to determine that distance. Our tool allows us this, by selecting the instance of our interest we obtain the Shapley values for that pattern that allow us to give an estimate of the contribution that each attribute brings to the group. The representation clearly distinguishes the individual contributions, for the pattern under consideration the attributes that cause the decrease in the accuracy value are in ascending order marital-status = "Married", hours-per-week => 45, and capital-gain = 0 .

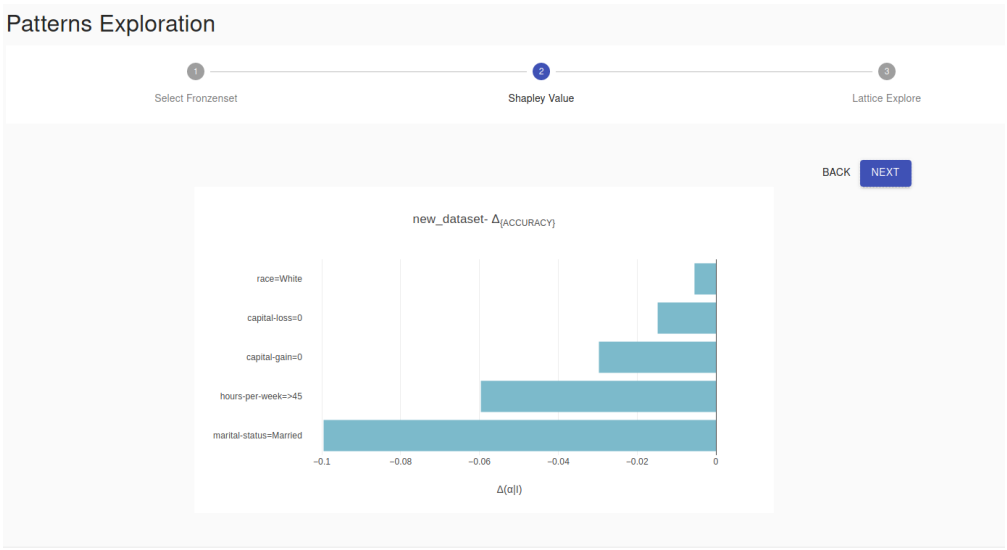


Figure 6.15: Shapley values for most discrepant pattern for Adult Dataset (Accuracy)

Different is the evaluation of the lattice search, we present the representation of the reticule without visualization of the regulatory items. The system checks if starting from a node and going down a level the group discrepancy value decreases, in which case it signals the regulator item. In the case of Accuracy values at each level the group discrepancy value decreases because they are negative values, we eliminate the representation of the regulatory items. The representation without adjustment serves in this case to give a better estimate of how the value of the group discrepancy increases with each addition of items.
To give an example, considering the local contributions examined thanks to the Shapley Value, let's take the three subgroups that will lead to a node above the threshold. As the Shapley values showed, the coalition between capital-gain = 0 and marital-status = "Married" leads to the highest group discrepancy value with respect to the other two subgroups, composed of the two separate attributes and the race = "White ", with which it will be combined. The race = "White" attribute has a minor contribution, but combined with the other attributes it increases the discrepancy value resulting in a node above the threshold.

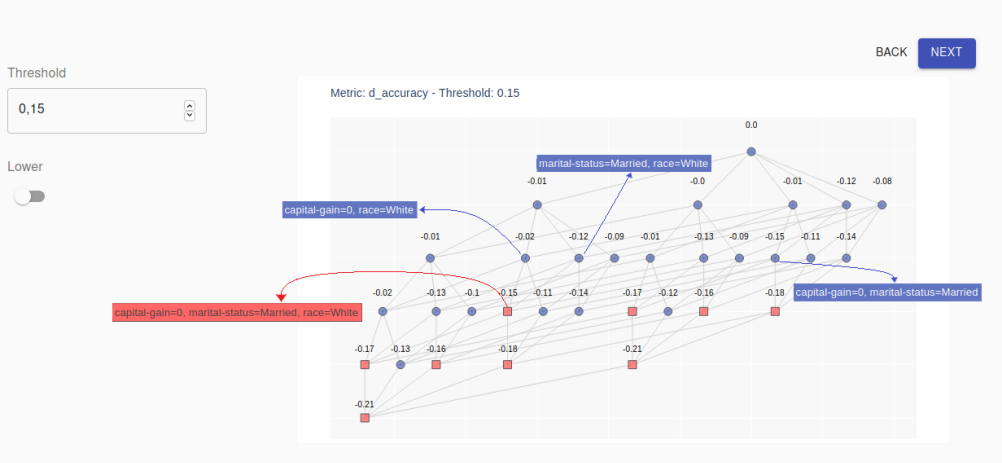


Figure 6.16: Lattice Search for most discrepant pattern for Adult Dataset (Accuracy)

Search for regulatory items

Even in the case of the search for regular items, the difference in Accuracy with respect to the previously evaluated metrics (6.1.1 and 6.1.2) must be emphasized. The adjustment takes place if the added item makes a positive contribution by increasing the group discrepancy from a negative value to a value close to 0. The example in the Figure 6.17 shows this concept well.
Consider the first instance of the table and the resulting Shapley Value, the addition of the item occupation = "Blue Collar" makes a positive contribution, even if not very large, which combined with the contribution of the hours-per-week attribute 40 dampens the negative effect deriving from the marital-status = "Married" attribute by increasing the group discrepancy value by a factor equal to 0.066 (from -0.141 to -0.074).

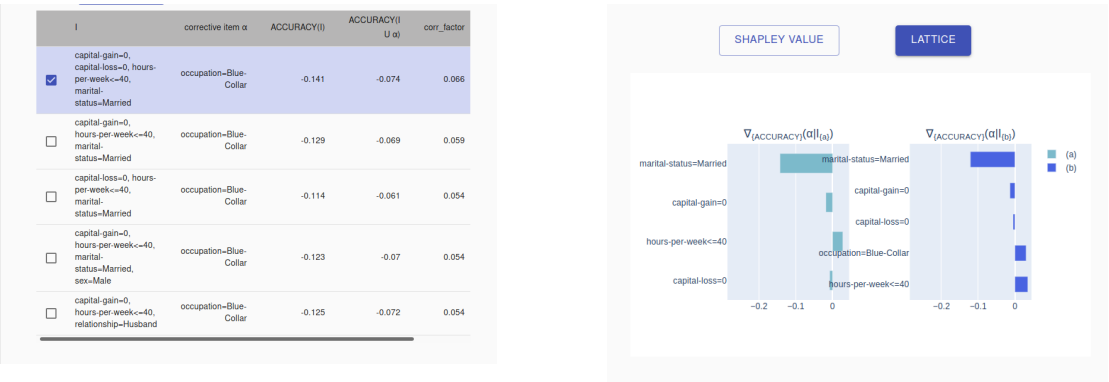


Figure 6.17: Regulatory item search for Accuracy (Adult Dataset)

We also evaluate the reticular search of the item set with the addition of the item occupation = "Blue Collar". The representation shows how the regulation affects a large slice of the lattice, obtaining most of the subgroups with group discrepancy values below the threshold (as can be seen, there is only one node with a value that exceeds the threshold equal to 0.17).

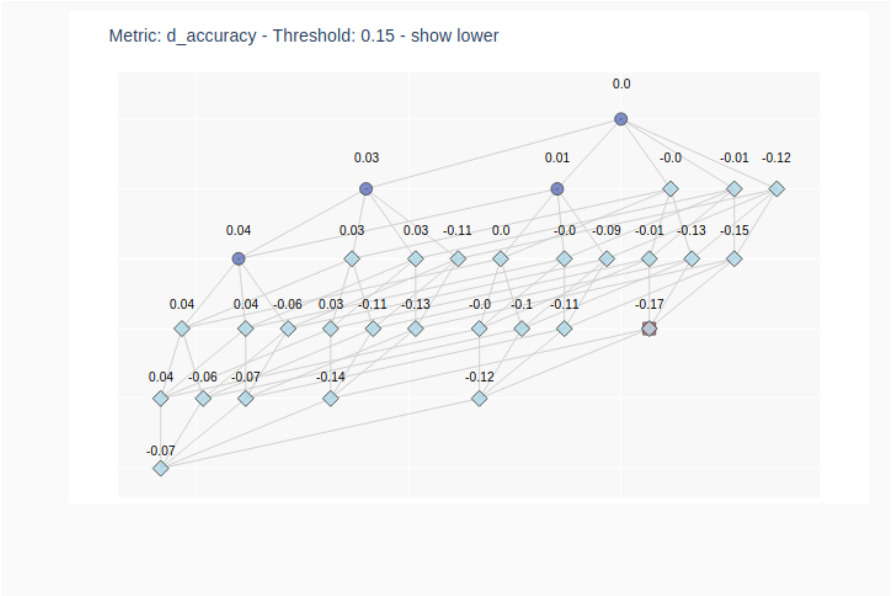


Figure 6.18: Lattice Search with the effect of the regulating item for Adult Dataset (Accuracy)

Global Evaluation

Finally, we evaluate the global results to determine the general contributions that the attributes give if added to an itemset and which coalitions determine an underperforming behavior of the model in the case of accuracy.
The representation of the Shapley values of all the attributes and their correlations places the attributes marital-status = "Married", hours-per-week => 45, and relationship = "Husband" on the podium. The correlation of these three attributes is the one with the greatest contribution to deviate the prediction from the correct value.

Figure 6.19: Global Discrepancy Group for every item for Accuracy (Adult Dataset)

6.2 Comparison with FairVis

In this section we will present a comparison with the FairVis [18] system. A first comparison will address the graphic differences that the two systems adopt. For the second comparison we will consider the presentation of results provided by the two systems.

6.2.1 Differences on the graphical interface level

In the first analysis, we can see a different set of interactive tools, as FairVis[18] proposes all the interaction operations on a single page, consequently, the user can make all choices in that window designed as shown in the figure 6.20.

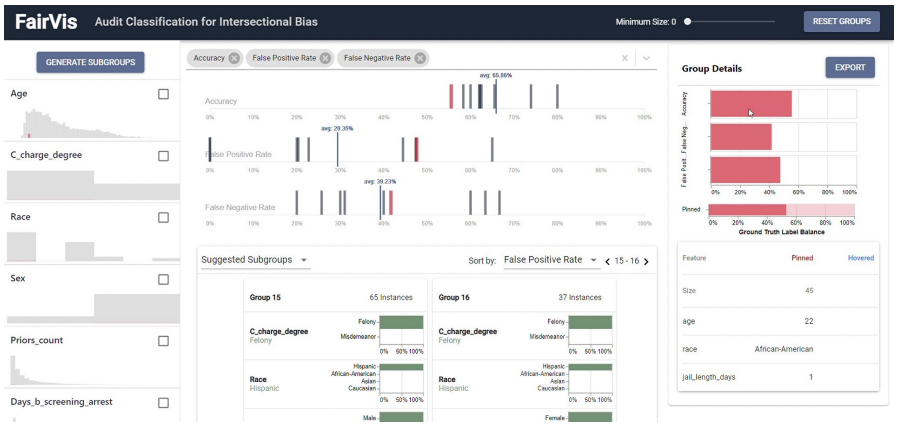


Figure 6.20: visualization of the tool proposed by FairVis[18]

In our proposal, on the other hand, we have created several views to isolate the passages considered important in the analysis and showing more clearly, for each passage, the results obtained at that specific moment.
The type of analysis that can be carried out in the two tools appears different despite the concepts on which they are based are similar.
In FairVis[18], the user can choose which attributes to select for the generation of sub groups, and the choice is aided by showing for each attribute value its distribution within the data set.
In our tool instead, the subgroups are automatically generated considering only the itemsets with a frequency greater than the support threshold entered by the user, in this way we obtain those groups of interest that have a certain relevance at the evaluation level giving a more general estimate of how the model behaves concerning the specific case not suitable for determining general behavior.
Once the groups with the inserted attributes have been generated, the evaluation

statistics chosen, FairVis[18] generates the result of the metric for each group by adding the display of the average, i.e. of the behavior concerning the whole set, to evaluate the distance of the subgroups concerning to the total; in our case, this distance, indicated in Chapter 4 as group discrepancy (definition 4.3.1), is the measure with which we go to estimate which patterns tend to be more significant for a wrong prediction.
Finally, FairVis[18], proposes to the user a set of subgroups statistically similar to those generated by the user, basing the similarity on the statistical divergence between the feature distributions to evaluate the differences in the impact on performance or to compare the results generated with subgroups more general with less functionality.
In our case, instead, we focus on the calculated group discrepancy value by analyzing the results obtained in more detail, we break down the subgroup into its elements obtaining, thanks to the Shapley values[44], the estimate of the contribution that each attribute has given for achieve that result, and also, let's evaluate the possibility of decreasing the distance concerning the overall by searching in an iterative process of adding an element to the item set, the items that have a regulatory effect on the distance, bringing the value closer to the general behavior.

6.2.2 Comparison presentation of results

To evaluate the different presentation of the results of the Fairvis system [18] compared to our tool, we take into consideration the evaluation of statistic Accuracy and we observe below the different representation of the results obtained.
We generate in Fairvis system [18], for convenience, subgroups similar, in terms of attribute value, to the subgroups that in our tool have the maximum and minimum Accuracy value. In Fairvis [18] it is possible to select for each attribute all the values available in the dataset in question, we have decided to take as example values those with a greater distribution within the dataset.

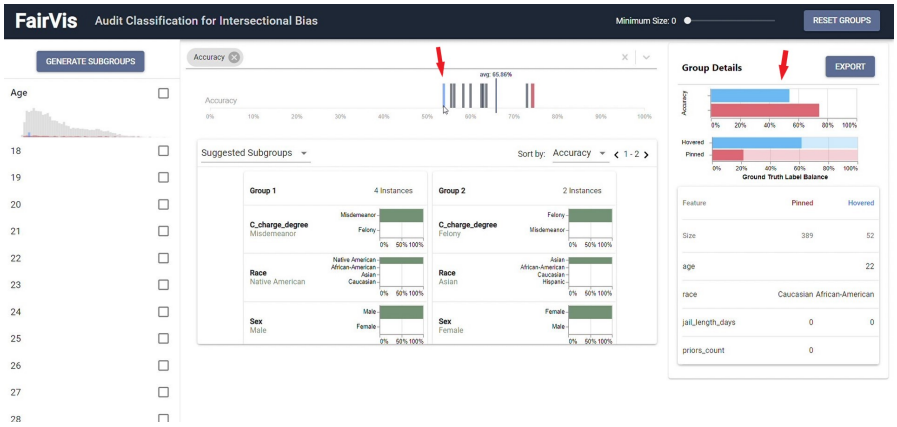
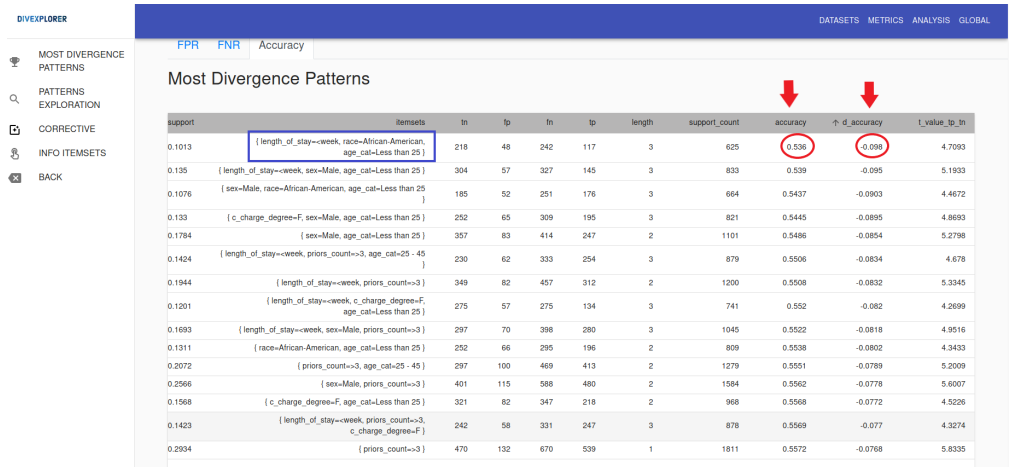


Figure 6.21: FairVis[18] accuracy evaluation: we generated the groups thanks to the panel on the left and we evaluated the two sub-groups that are at the extremes compared to the average of the overall; on the left the bar graph shows in red the accuracy of the subgroup to the right of the average, in blue the accuracy of the subgroup to the left of the average.

From the representation given to us, we can evaluate for each subgroup its distance from the general behavior indicated by the long bar with the number above; at this point let us examine the two extreme cases as shown by the arrows. FairVis[18] allows the selection of a group to keep still in the comparison and the possibility of indicating the other group by simply passing the mouse over the various bars generated, indicating the stationary group in red and the mobile group in blue.
From the analysis shown on the right we see a bar graph with different colors to show the relative accuracy value for the two groups and below the detailed information relating to the groups examined. FairVis[18] aims to evaluate the fairness of a model, we cannot fail to notice in this case that the two groups taken into consideration, for the same days in prison, have a very significant distance from each other,but not being able to quantify the influence of each attribute, it is difficult to determine if the cause of this detachment can be traced back to another comparable attribute, that is race.



support	itemsets	tn	fp	fn	tp	length	support_count	accuracy	↑ d_accuracy	t_value_tp_tn
0.1013	{ length_of_stay=<week, race=African-American, age_cat=Less than 25 }	218	48	242	117	3	625	0.536	-0.098	4.7093
0.135	{ length_of_stay=<week, sex=Male, age_cat=Less than 25 }	304	57	327	145	3	833	0.539	-0.095	5.1933
0.1076	{ sex=Male, race=African-American, age_cat=Less than 25 }	185	52	251	176	3	664	0.5437	-0.0903	4.4672
0.133	{ c_charge_degree=F, sex=Male, age_cat=Less than 25 }	252	65	309	195	3	821	0.5445	-0.0895	4.8693
0.1784	{ sex=Male, age_cat=Less than 25 }	357	83	414	247	2	1101	0.5486	-0.0854	5.2798
0.1424	{ length_of_stay=<week, priors_count=>3, age_cat=25 - 45 }	230	62	333	254	3	879	0.5506	-0.0834	4.678
0.1944	{ length_of_stay=<week, priors_count=>3 }	349	82	457	312	2	1200	0.5508	-0.0832	5.3345
0.1201	{ length_of_stay=<week, c_charge_degree=F, age_cat=Less than 25 }	275	57	275	134	3	741	0.552	-0.082	4.2699
0.1693	{ length_of_stay=<week, sex=Male, priors_count=>3 }	297	70	398	280	3	1045	0.5522	-0.0818	4.9516
0.1311	{ race=African-American, age_cat=Less than 25 }	252	66	295	196	2	809	0.5538	-0.0802	4.3433
0.2072	{ priors_count=>3, age_cat=25 - 45 }	297	100	469	413	2	1279	0.5551	-0.0789	5.2009
0.2566	{ sex=Male, priors_count=>3 }	401	115	588	480	2	1584	0.5562	-0.0778	5.6007
0.1568	{ c_charge_degree=F, age_cat=Less than 25 }	321	82	347	218	2	968	0.5568	-0.0772	4.5226
0.1423	{ length_of_stay=<week, priors_count=>3, c_charge_degree=F }	242	58	331	247	3	878	0.5569	-0.077	4.3274
0.2934	{ priors_count=>3 }	470	132	670	539	1	1811	0.5572	-0.0768	5.8335

Figure 6.22: Accuracy evaluation in our interactive tool: we have the table with all the subgroups automatically generated by the Frequent Pattern Mining algorithm; it is possible to sort the columns of interest in our case we evaluate the lowest accuracy values to verify how far they are from the overall.

Our model, on the other hand, proposes a table of frequent subgroups in which we can evaluate all the characteristics of a given instance, we present its support value, the composition of the item set, the confusion matrix, and all the measures relating to the statistic selected among which is the group discrepancy value (the two red circles in the image show the lowest accuracy value and its corresponding group discrepancy value). We want to better analyze the pattern with the lowest discrepancy value, thanks to the navigation bar on the left we can move to the second view.

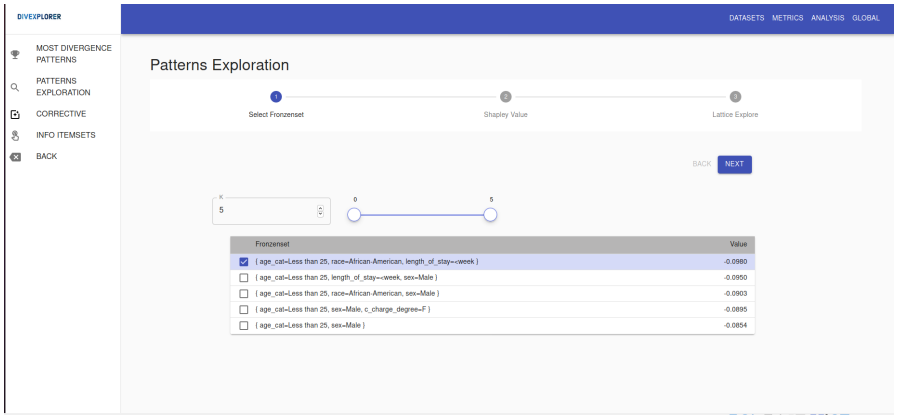


Figure 6.23: Pattern Exploration: this view allows the user to evaluate, for each generated subgroup, how the group discrepancy value is produced; selecting the row of interest can display a series of details such as the lattice graph.

We select the row corresponding to the group discrepancy value we are interested in and click next to obtain the corresponding Shapley Value[44].

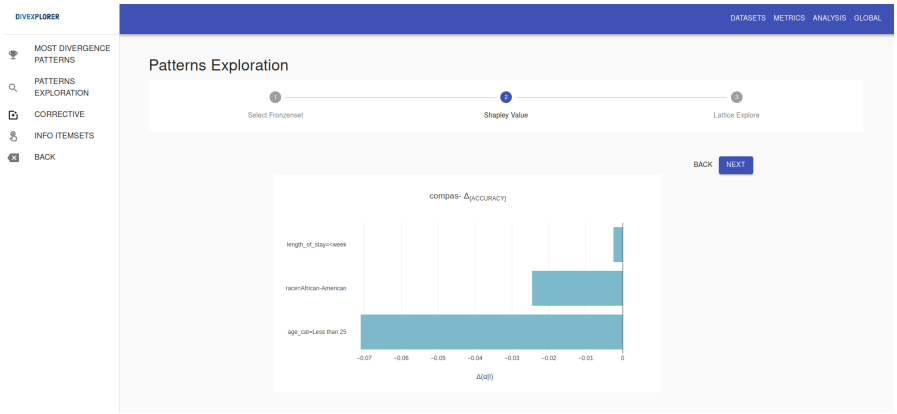


Figure 6.24: Pattern Exploration Shapley Value: selected the instance to be evaluated by pressing the next button, the user will see the bar graph that represents the Shapley values for the elements of the item set.

From the Shapley Value[44], we can see that the attribute that contributed most to obtaining such a low accuracy value was the age followed immediately after by the breed. We perform the same operations for the item set with the highest accuracy value to make a complete comparison, select the corresponding row and generate the corresponding graphic for the Shapley Value.

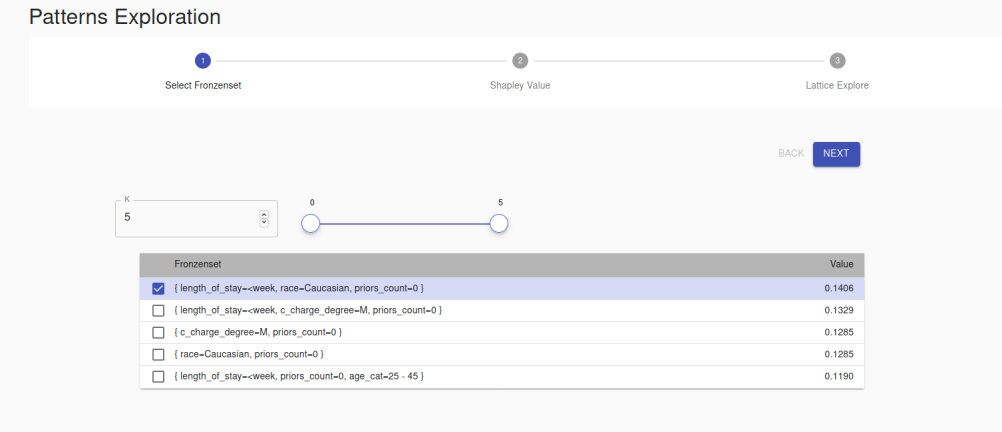


Figure 6.25: Pattern Exploration higher accuracy values.

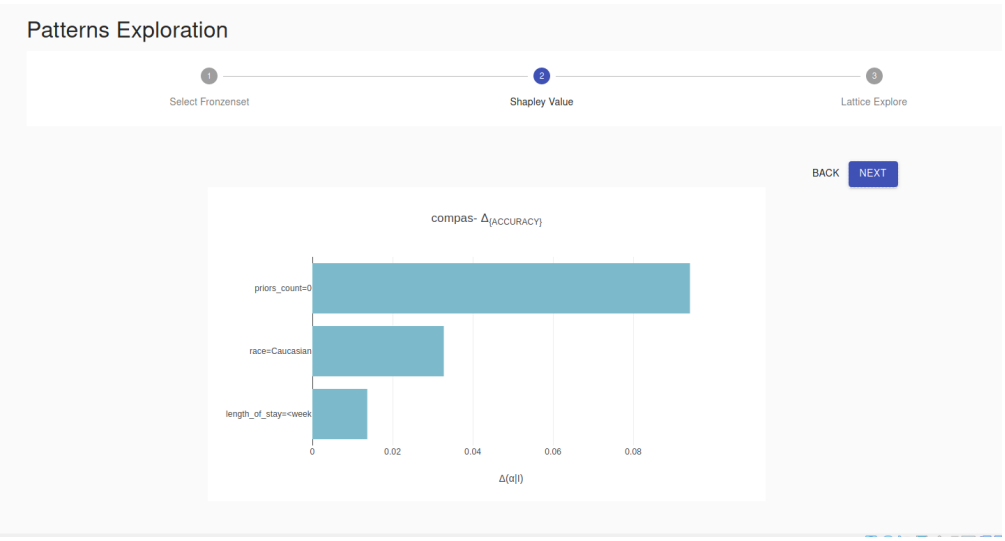


Figure 6.26: Pattern Exploration Shapley Value higher accuracy values.

In this case, we can see that the factor that has contributed most to having a high accuracy value is never having been in prison followed by being Caucasian; we also propose the lattice graph to explore how the group discrepancy varies in the generation of the lattice of the selected item.

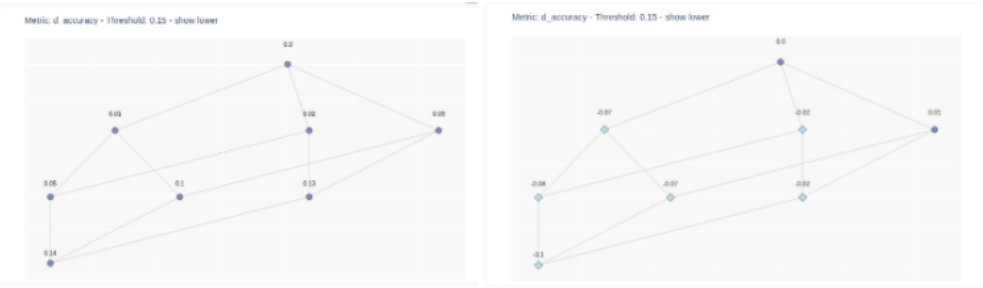


Figure 6.27: Pattern Exploration Lattice Graph:on the left the lattice graph corresponding to the instance with the highest accuracy value; on the right the lattice graph corresponding to the instance with the value of less than accuracy.

Thanks to our research of the elements that regulate the group discrepancy, taking as an example the last case examined, the priors _count = 0 attribute will be just one of these. As we can see in the Figure 6.28, we see that the accuracy value goes from -0.039 to 0.015 by adjusting the group discrepancy by a factor of 0.023. Furthermore, the Shapley Value[44] on the right makes the concept even clearer while race = African-American and sex = Male tend to lower the result priors _count = 0 it contributes positively by correcting the final value.

	I	corrective item α	ACCURACY(I)	ACCURACY(I U α)	corr_factor	t_corr
☐	#prior=0, sex=Male	race=Afr-Am	0.077	0.015	0.061	2.8
☐	stay<week, #prior=0	race=Afr-Am	0.101	0.043	0.058	3
☐	#prior=0	race=Afr-Am	0.093	0.04	0.053	2.8
☐	charge=M, stay<week	#prior=[1,3]	0.045	-0.007	0.037	2.4
☐	charge=M, stay<week	race=Afr-Am	0.045	-0.009	0.036	2.7
☑	race=Afr-Am, sex=Male	#prior=0	-0.039	0.015	0.023	2.5
☐	charge=M	race=Afr-Am	0.031	-0.014	0.017	2.4

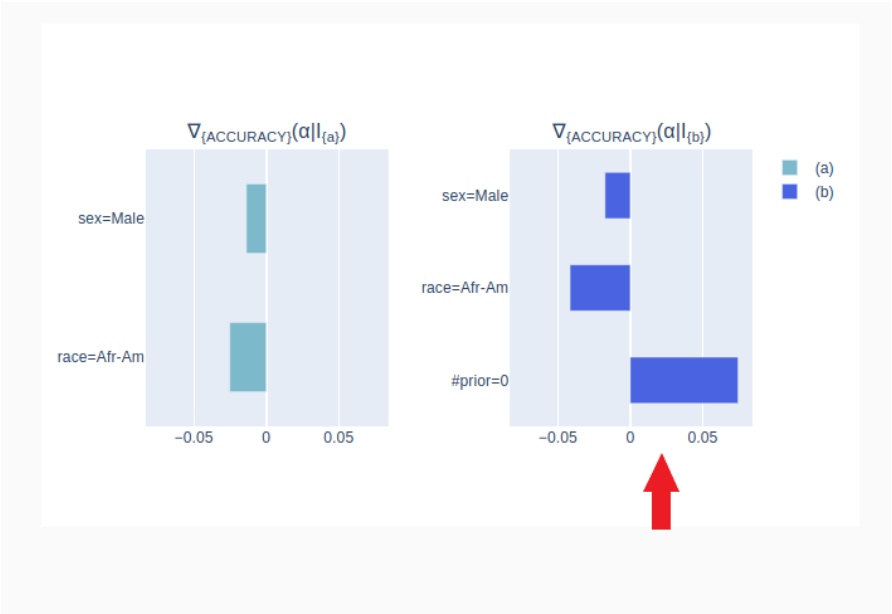


Figure 6.28: Evaluation of items that decrease the distance value

6.3 Comparison with Aequitas

We know that the Aequitas system (2.1.5) is a toolkit for discovering the presence of bias or lack of fairness of Machine Learning models. The toolkit is very simple to use, the user only has to load the dataset to be analyzed and set the bias metrics and protected attributes of interest. The groups generated by the toolkit are formed by all the entities that share the same attribute value, e.g. sex = "Male".
We compare our tool with the Aequitas toolkit on the basis of the evaluation and presentation of the results.

6.3.1 Comparison of results presentation

We propose the visualization of the results of the Notebook provided to carry out the analysis of the COMPAS dataset with the approach proposed by Aequitas. The analysis involves setting the evaluation metrics, we choose the False Positive Rate (FPR). As attributes of interest we consider age, race and sex which represent the sensitive attributes considering the context of analysis (the prediction of the recidivism of a defendant) and we evaluate, with respect to these attributes, how the model behaves.

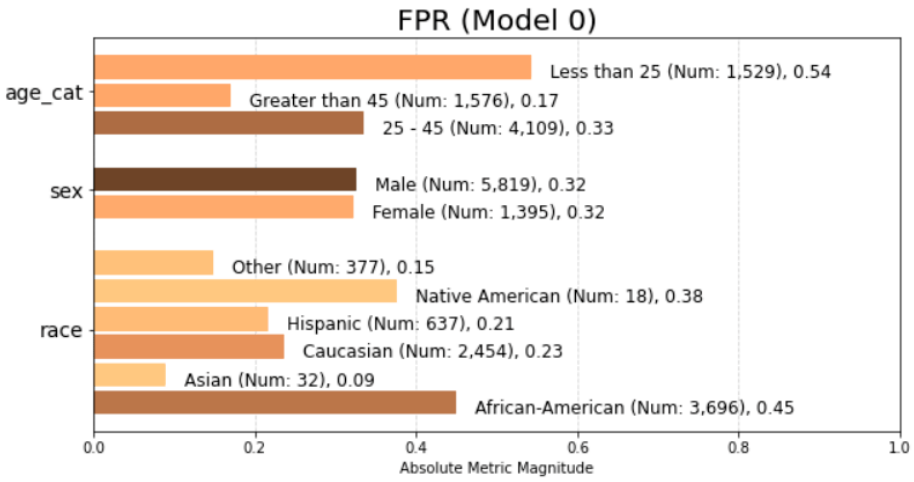


Figure 6.29: Evaluation of the False Positive Rate for sensitive characteristics (Aequitas [106]).

From the visualization of the Figure 6.29, we can immediately determine which attribute values carry a greater probability of being classified as false positives. The representation shows for each attribute inserted a bar graph for each valure th The results under examination lead to the conclusion that attribute can take,

identifying the difference in value with different gradations of the orange color. For example, from the graph obtained we can evaluate that men and women are equally likely, but African American people under the age of 25 have a much higher false positive rate (0.54 and 0.45 for age _cat and race respectively) when compared with Asian people over the age of 45. Always using the Aequitas toolkit we search for biases for age and race attributes. In this case, the toolkit allows the comparison of all attribute values with respect to one considered as a reference and computes the disparity for each reference/value pair. We consider the age _cat attribute and as a reference the value 25-45 (value with the average result).

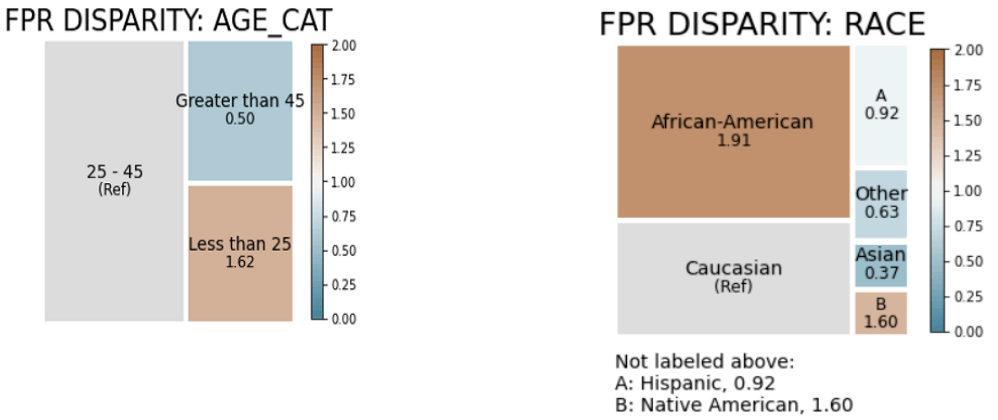


Figure 6.30: Visualizing disparities between groups in the single attribute age and race for FPR (Aequitas [106])

The display in the Figure 6.30 offers a clearer idea of the present bias, in fact, taking the average value as a reference, the result of the disparities is decidedly significant. For the False Positive Rate we have for age _cat = “Less than 45” a disparity equal to 1.62.
By performing the same operations on the race attribute and considering race = "Caucasian" as a reference, we obtain incredible results. We see that the disparity value for the cases of false positives is equal to 1.91 for African-Americans and 1.60 for Native-Americans, it is a very high disparity considering that the disparity of the reference attribute is equal to 1.
Finally we evaluate the fairness of the model for the sensitive attributes that we have decided to analyze. Fairness is evaluated by taking into consideration the levels of disparity calculated during the search for the Bias, and using the "80% rule" [19] that is, considering a threshold th = 0.8 (equivalent to 80%), a certain group is considered fair if its disparity value is between 0.8 and 1.25 (corresponding to $\frac{1}{th}$).

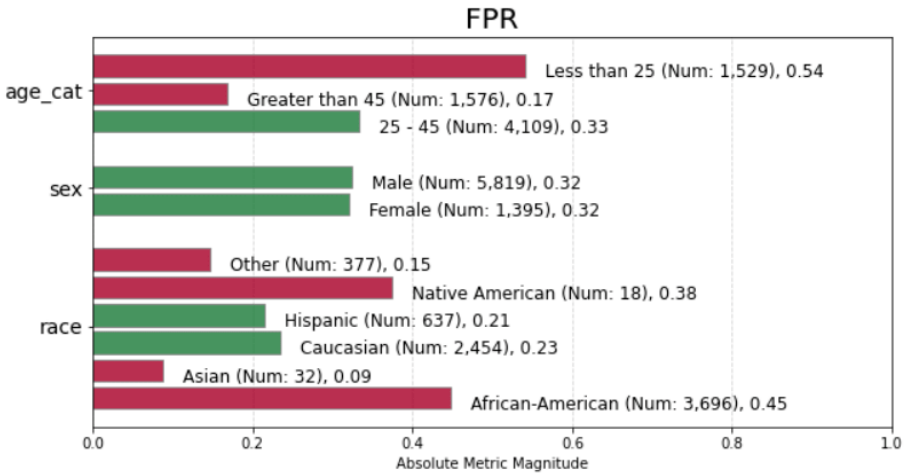


Figure 6.31: Visualizing fairness of a single absolute group metric across all population groups (Aequitas [106])

Also in the case of fairness the reference values must be set, in our case we have race = "Caucasian", sex = "Male" and age _cat = "25 - 45". The representation in the Figure, very simple to read, shows a bar graph where green indicates fair group, red indicates unfair group. We note that as fair groups with respect to the reference we have only sex = "Female" and race = "Hispanic", all other groups are considered unfair with respect to the reference, there is no statistical parity.
As in the case of Bias, it is possible to visualize the disparity about fairness for a single attribute and metric. We show the results for the "race" attribute for the False Positive Rate statistic. We note how the contrast of colors leaves no doubt to the user, almost all the blocks are colored in red, a sign that cannot be considered statistically equal to the gray colored reference. For more details see [107]

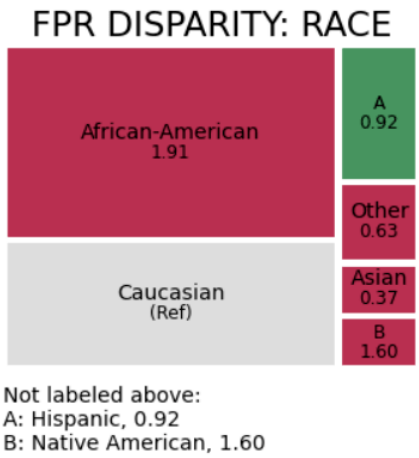


Figure 6.32: Visualizing fairness between groups in a single attribute race for all calculated disparity metrics (Aequitas [106])

Let's now consider the same conditions but developed within our tool with respect the classifier that we have used. Our tool does not yet have the choice of individual attribute values to be analyzed but we provide at the end of the analysis the global results that can be an aspect of comparison with the Aequitas system [107] in evaluating the fairness of the model.
Recall that the global results provide for each statistic examined, a representation of the Shapley values of each attribute value, first of all showing the individual contributions that each feature makes in addition to other itemsets. Furthermore, this type of evaluation offers the possibility to determine which combinations of attributes cause an increase or a decrease of the group discrepancy value for each metric (definition 4.3.1). It is a more detailed type of assessment with which to verify the fairness and bias present within a Machine Learning model.

Figure 6.33: Representation global Shapley values for FPR and FNR

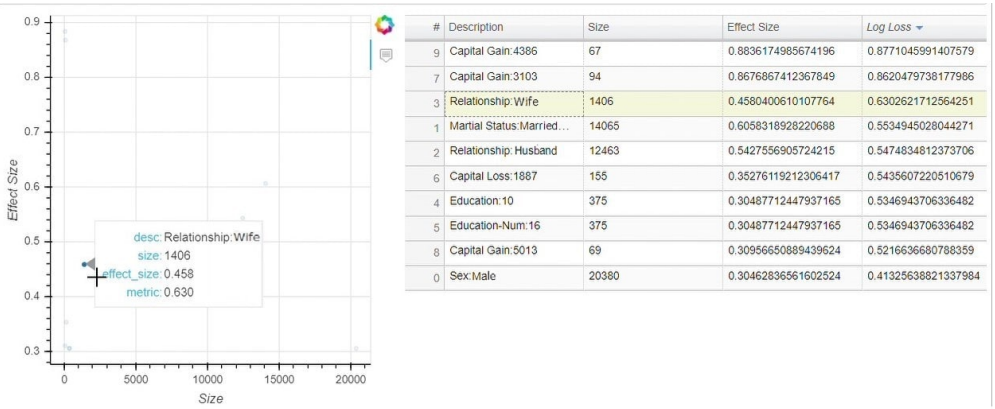
We present the results obtained for the COMPAS dataset considering the same metric analysed previously, False Positive Rate.
The global representation provides clear results to evaluate fairness by adding also the quantification of the discrepancy value for each attribute value, so we can determine not only which characteristics can be considered more critical at the level of incorrect prediction but we also visualize the measurement of the effect which certain characteristic contributes to the final result.

6.4 Comparison with Slice Finder

The Slice Finder system (2.1.1) allows the identification of problematic sections in which the model's performance is not correct. The first consideration to be addressed concerns the setting of the solution. The Slice Finder system determines the significant difference in model performance metrics in terms of a section's log loss compared to its counterpart. Furthermore, to determine if the loss is significant, they consider the size of the effect of the difference. We discuss the presentation of the results and the interactivess of Slice Finder (2.1.1) on the Adult dataset.

6.4.1 Comparison of results presentation with Adult Dataset

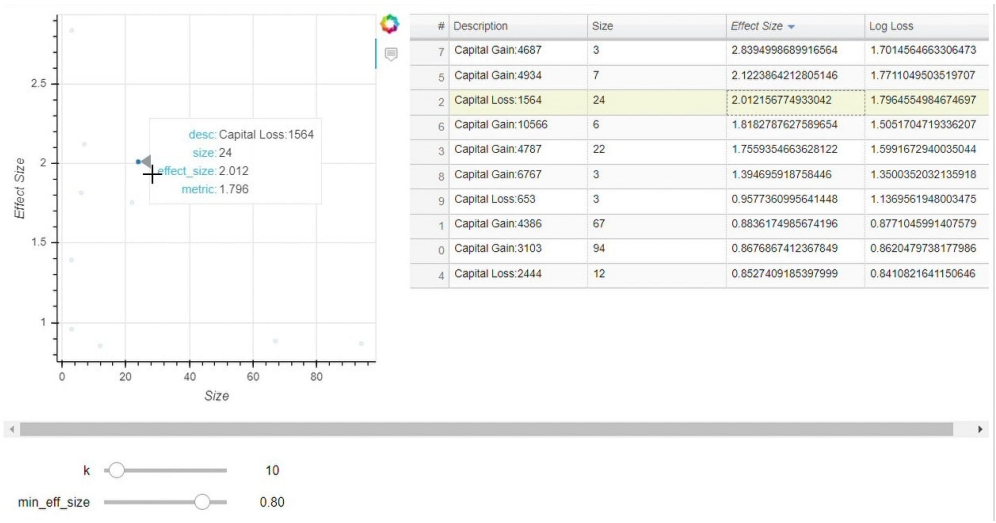
The example in the Figure 6.34 shows the interactive view offered by the Slice Finder system. The user selects a K number and an effect size threshold and obtains the top-k sections that can be explored with an effect size smaller than the set threshold. It is possible to sort the table according to the column of your interest, in this case we sort according to the difference of the logarithmic loss and select the row relating to relationship = "Wife" to display the position of the attribute value in the scatter chart. By placing the mouse on the corresponding point, you can view a summary of the information of the selected item.



#	Description	Size	Effect Size	Log Loss
9	Capital Gain:4386	67	0.8836174985674196	0.8771045991407579
7	Capital Gain:3103	94	0.8676867412367849	0.8620479738177986
3	Relationship:Wife	1406	0.4580400610107764	0.6302621712564251
1	Martial Status:Married...	14065	0.6058318928220688	0.5534945028044271
2	Relationship:Husband	12463	0.5427556905724215	0.5474834812373706
6	Capital Loss:1887	155	0.35276119212306417	0.5435607220510679
4	Education:10	375	0.30487712447937165	0.5346943706336482
5	Education-Num:16	375	0.30487712447937165	0.5346943706336482
8	Capital Gain:5013	69	0.30956650889439624	0.5216636680788359
0	Sex:Male	20380	0.30462836561602524	0.41325638821337984

Figure 6.34: Evaluation of problem sections for the Slice Finder system [108]

A section is defined as problematic if the difference in loss is statistically significant and if the size of the effect of the difference is large enough (from 0.8 upwards), we try to sort according to the size of the effect. We obtain as top-k the sections relevant to the capital-gain and capital-loss attributes.
Our approach evaluates the problematic subgroups through the difference in performance with the entire dataset (definition of group discrepancy 4.3.1). The



#	Description	Size	Effect Size	Log Loss
7	Capital Gain:4687	3	2.8394998689916564	1.7014564663306473
5	Capital Gain:4934	7	2.1223864212805146	1.7711049503519707
2	Capital Loss:1564	24	2.012156774933042	1.7964554984674697
6	Capital Gain:10566	6	1.8182787627589654	1.5051704719336207
3	Capital Gain:4787	22	1.7559354663628122	1.5991672940035044
8	Capital Gain:6767	3	1.394695918758446	1.3500352032135918
9	Capital Loss:653	3	0.9577360995641448	1.1369561948003475
1	Capital Gain:4386	67	0.8836174985674196	0.8771045991407579
0	Capital Gain:3103	94	0.8676867412367849	0.8620479738177986
4	Capital Loss:2444	12	0.8527409185397999	0.8410821641150646

Figure 6.35: Evaluation of problem sections with max effect size for the Slice Finder system [108]

selection of the subgroup to be analyzed allows a more detailed exploration of the individual components of the pattern and a more intuitive view of the effect that each attribute value has on the final result. We also provide to the user the search for regulatory items that tend to control the contribution of the most critical functionalities, bringing the result closer to the correct value. Finally, we provide the global results that allow greater clarity in the evaluation of the overall functionality of the dataset in question. The detailed analysis of the Adult dataset with our tool is described in Section 6.1.

Chapter 7

Conclusion

There are contexts in which the inclusion of machine learning systems as a method of choice brings to attention an intrinsic problem relating to the possible prejudices that the system could verify during the classification when the result of the prediction influences the human being.
In this thesis, we have analyzed a new data exploration approach for classification analysis that is based on the concept of group discrepancy. The group discrepancy is an estimate of the statistical distance of a given pattern with respect to the entire dataset, giving a first idea of the behavior of the model in representative subgroups of the dataset that can be critical. We explored the group discrepancy on a more detailed level by providing Shapley values for the individual features that make up the itemset to assess which attributes are most responsible for deviating from general behavior. We also provide the lattice research of the pattern in which to evaluate every possible combination of attributes and the contribution it has on the group discrepancy.
We provide the possibility to search for items with a regulatory effect by evaluating, as in a coalition game, the difference in group discrepancy of the pattern with and without the added item, displaying in detail the local contributions of each attribute. Finally, we propose the definition of global group discrepancy for the evaluation of the influence of one functionality in correlation with the others. From these results, we are able to determine which groups cause inaccuracy of the prediction and we can quantify for each attribute its group discrepancy value.
The thesis also wants to underline the importance of bringing the user closer to exploring the approach through a more interactive system. User involvement becomes an essential aspect to consider to increase the diffusion of new approaches for machine learning systems and to improve the understanding of the method used.
We present a low-level interactive approach based on carrying out operations that require user action in the context of Jupyter Notebooks, making the analysis process

more dynamic and allowing the realization of different hypotheses without having to search within the code the edit point. The creation of an interactive notebook allows a more professional user experience, requires an understanding of the code but speeds up editing operations by updating the modified result online. The creation of an interactive notebook allows a more professional user experience, requires an understanding of the code but speeds up editing operations by updating the modified result online.
We also propose a high-level approach with the creation of an interactive tool in which the user remains extraneous to the implementation of the method and is completely immersed in the visual analysis of the results he can obtain. The application consists of several views in which the user can carry out a part of the analysis through the selection of parameters, the interactive display of explanatory graphs or the manipulation of data tables. Attention to detail, the fluidity of the executive process, and ease of use are the characteristics that make our tool a valid proposal for a new way of undertaking a process of optimization or analysis of machine learning systems. We focus on the definition of usability, the key principle of human-computer interaction, to show that the optimization of systems can also be aimed at the user side and not just the code side, the direct involvement of the user in a research context can simplify the understanding process and allow faster development of solutions for system anomalies.

7.1 Future Work

Human-computer interaction is a growing concept that still has to outline a precise application protocol even if it is strongly requested in the new emerging systems used by human, such as those of machine learning. Future work provides the optimization of the presented interactive tool, increasing the possibilities offered to the user at the analysis level, for example by exploring the possible extensions of the concept of group discrepancy. A design improvement is also provided, allowing the customization of the graphics displayed both at a stylistic and structural level. Another possible optimization concerns the expansion of the tool distribution platforms considering also other operating systems (Windows, Android). In this way, we can increase the versatility of the presented work allowing its use on any device.
Future work will focus on the creation of a complete application in all its features, abandoning the definition of demo in favor of a more full-bodied tool, in order to be able to build analysis and not just verify the results.

Bibliography

[1] Tom M.Mitchell. *The Discipline of Machine Learning*. Machine Learning Department technical report CMU-ML-06-108. Carnegie Mellon University, July 2006 (cit. on pp. 1, 2).

[2] Mohri Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. «Foundations of machine learning». In: MIT press. 2018 (cit. on p. 1).

[3] Bonaccorso Giuseppe. «Machine learning algorithms». In: Packt Publishing Ltd, 2017 (cit. on p. 1).

[4] Brynjolfsson Erik and Mitchell Tom. «What Can Machine Learning Do? Workforce Implications.» In: Science (American Association for the Advancement of Science) 358.6370 (2017): 1530-534. Web. (cit. on p. 1).

[5] Gori. «Machine Learning». In: Morgan Kaufmann, 2018. Web (cit. on p. 1).

[6] Mohammed Mohssen, Khan Muhammad Badruddin, and Bashier Eihab Bashier Mohammed. «Machine Learning». In: CRC, 2016. Web (cit. on p. 1).

[7] Abbeel Pieter and other. «An application of reinforcement learning to aerobatic helicopter flight». In: *Advances in neural information processing systems*. 19 (2007): 1 (cit. on p. 2).

[8] Ng Andrew Y. and other. «Autonomous inverted helicopter flight via reinforcement learning». In: *Experimental robotics IX*. Springer, Berlin, Heidelberg, 2006. 363-372 (cit. on p. 2).

[9] Abbeel Pieter, Adam Coates, and Andrew Y. Ng. «Autonomous helicopter aerobatics through apprenticeship learning». In: *The International Journal of Robotics Research*. 29.13 (2010): 1608-1639 (cit. on p. 2).

[10] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. *Machine Bias*. 2016. URL: `https://www.propublica.org/article/machine-bias-risk-assesments-in-criminal-sentencing` (cit. on pp. 2, 13).

[11] S. Barocas and A. D. Selbst. «Big data's disparate impact». In: Cal. L. Rev., 104:671, 2016 (cit. on p. 2).

[12] A. Chouldechova. «Fair prediction with disparate impact: A study of bias in recidivism prediction instruments». In: *Big data, 5(2):153–163, 2017* (cit. on p. 3).

[13] Eliana Pastor, Luca de Alfaro, and Elena Baralis. «Looking for Trouble: Analyzing Classifier Behavior via Pattern Divergence». In: *In Proceedings of the 2021 International Conference on Management of Data (SIGMOD '21), June 20–25, 2021, Virtual Event, China. ACM, New York, NY, USA, 13 pages.* (2021). URL: https://doi.org/10.1145/3448016.3457284 (cit. on pp. 3, 19, 22).

[14] E. Strumbelj and I. Kononenko. «An efficient explanation of individual classifications using game theory». In: *Journal of Machine Learning Research* 11 (2010) (cit. on pp. 4, 18).

[15] Y.Chung, T.Kraska, K.H.Tae N.Polyzotis, and S.E.Whang. «Slice Finder: Automated Data Slicing for Model Validation». In: *IEEE 35th International Conference on Data Engineering(ICDE).* 2019, pp. 1550–1553 (cit. on pp. 4, 5, 10).

[16] Dezhi Fang Minsuk Kahng and Duen Horng Chau. «Visual exploration of machine learning results using data cube analysis». In: *In Proceedings of the Workshop on Human-In-the-Loop Data Analytics.* 2016, pp. 1–6 (cit. on pp. 5, 10).

[17] R.K.E.Bellamy et al. «AI Fairness 360:An extensible toolkit for detecting and mitigating algorithmic bias». In: *IBM Journal of Research and Development* 63 (2019), pp. 4–5 (cit. on pp. 5, 10, 11).

[18] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Jamie Morgenstern Minsuk Kahng, and Duen Horng Chau. «FairVis: Visual analytics for discovering intersectional bias in machine learning». In: *IEEE Conference on Visual Analytics Science and Technology (VAST).* 2019, pp. 46–56. URL: https://poloclub.github.io/FairVis/ (cit. on pp. 6, 9, 11, 13, 71, 86–88).

[19] Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. «Aequitas: A Bias and Fairness Audit Toolkit». In: (2018). arXiv preprint arXiv:1811.05577 (cit. on pp. 6, 11, 95).

[20] Amershi, S., Cakmak, M., Knox, W. B., Kulesza, and T. «Power to the people: The role of humans in interactive machine learning». In: *AI Magazine* (). in press (cit. on pp. 7, 31).

[21] Arvind Narayanan. «Translation tutorial: 21 fairness definitions and their politics». In: *In Conference on Fairness, Accountability, and Transparency.* 2018 (cit. on pp. 7, 12).

[22] D.A.Keim, T.Munzner, F.Rossi, and M.Verleysen. «Bridging Information Visualization with Machine Learning». In: *Dagstuhl Seminar 15101*. Vol. 5 (3). Dagstuhl Reports, 2015, pp. 1–27 (cit. on p. 7).

[23] Josua Krause, Adam Perer, and Kenney Ng. «Interacting with predictions: Visual inspection of black-box machine learning models». In: *In Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 5686– 5697 (cit. on p. 8).

[24] P.Tamagnini, J.Krause, A.Dasgupta, and E. Bertini. «Interpreting blackbox classifiers using instance-level visual explanations». In: *in Proc. 2nd Workshop Hum.-Loop Data Anal.* Art. no. 2017 (cit. on p. 8).

[25] A. Kapoor, B.Lee, D.Tan, and E.Horvitz. «Performance and preferences: Interactive refinement of machine learning procedures». In: *In Proc. of AAAI 2012*. 2012 (cit. on p. 8).

[26] D.Ren, S.Amershi, B.Lee, J.Suh, and J.D.Williams. «Squares: Supporting interactive performance analysis for multiclass classifiers». In: *IEEE Transactions on Visualization and Computer Graphics*. Vol. 23(1):61–70. 2017 (cit. on p. 8).

[27] S.K.Badam, N.Elmqvist, and J.-D.Fekete. «Steering the craft: Ui elements and visualizations for supporting progressive visual analytics». In: *Comput. Graph.* Vol. 36(3). Forum, 2017 (cit. on p. 9).

[28] Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan 2011a. «Effective end-user interaction with machine learning». In: *In Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*. AAAI Press, 2011, pp. 1529–1532 (cit. on p. 9).

[29] Han Jiawei, Pei Jian, and Kamber Micheline. «Data Mining: Concepts and Techniques». In: *In Handbook on information technologies for education and training*. Morgan Kaufmann, 2011. The Morgan Kaufmann Ser. in Data Management Systems. Web. (cit. on pp. 11, 20).

[30] Dulli Susi, Biscari P, Furini Sara, and Peron Edmondo. «Data Mining». In: Springer Milan, 2009. Web. (cit. on pp. 11, 20).

[31] Aggarwal Charu C. «Data Classification». In: Vol. 35. Philadelphia, PA: Chapman and Hall/CRC, 2015. Chapman Hall/CRC Data Mining and Knowledge Discovery Ser. Web. (cit. on p. 11).

[32] Zliobaite Indre. «On the relation between accuracy and fairness in binary classification». In: arXiv preprint arXiv:1505.05723 (2015). (cit. on p. 11).

[33] Barocas Solon, Moritz Hardt, and Arvind Narayanan. «Fairness in machine learning». In: Nips tutorial 1 (2017): 2. (cit. on p. 11).

[34] M.J.Kusner, J.R.Loftus, C.Russell, and R. Silva. «Counterfactual fairness». In: (2018). arXiv preprint arXiv:1703.06856v3 (cit. on pp. 11, 12).

[35] S. Barocas and A. D. Selbst. «Big data's disparate impact». In: *Cal. L. Rev.* Vol. 104:671. 2016 (cit. on p. 11).

[36] Alexandra Chouldechova. «Fair prediction with disparate impact: a study of bias in recidivism prediction instruments». In: *Big Data*. Vol. 2:153–163. 2017 (cit. on pp. 11, 13).

[37] Nexa Center for Internet Society.2018. «Fairness e Machine Learning Il concetto di equità e relative formalizzazioni nel campo dell'apprendimento automatico». In: (). URL: `https://nexa.polito.it/nexacenterfiles/Articolo%20TIM.pdf` (cit. on p. 12).

[38] Verma Sahil and Julia Rubin. «Fairness definitions explained». In: *ieee/acm international workshop on software fairness (fairware). IEEE, 2018*. 2018 (cit. on pp. 12, 21).

[39] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. «On the (im) possibility of fairness». In: (2016). arXiv preprint arXiv:1609.07236 (cit. on p. 13).

[40] J. M. Kleinberg, S. Mullainathan, and M. Raghavan. «Inherent trade-offs in the fair determination of risk scores». In: *ITCS*. 2017 (cit. on p. 13).

[41] D. Xu, S. Yuan, L. Zhang, and X. Wu. «Fairgan: Fairness-aware generative adversarial networks». In: *IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 570–575 (cit. on p. 13).

[42] A. Beutel, J. Chen, T. Doshi, H. Qian, A. Woodruff, C. Luu, P. Kreitmann, J. Bischof, and E. H. Chi. «Putting fairness principles into practice: Challenges, metrics, and improvements». In: *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*. AIES, 2019 (cit. on p. 13).

[43] M. Hardt, E. Price, N. Srebro, et al. «Equality of opportunity in supervised learning». In: *Advances in neural information processing systems*. 2016, pp. 3315–3323 (cit. on p. 13).

[44] Lloyd S. Shapley. «A Value for n-person Games». In: *volume II of Contributions to the Theory of Games*. Princeton University Press, 1953 (cit. on pp. 14, 15, 17, 18, 24–26, 87, 90–92).

[45] Morgenstern Oskar and John Von Neumann. «Theory of games and economic behavior». In: Princeton University Press, 1953. (cit. on p. 14).

[46] Stefano Moretti and Fioravante Patrone. «Transversality of the shapley value». In: TOP, 16(1):1–41. 2008 (cit. on pp. 17, 18).

[47] Alon Keinan, Ben Sandbank, Claus C. Hilgetag, Isaac Meilijson, and Eytan Ruppin. «Fair attribution of functional contribution in artificial and biological networks». In: *Neural Computation*. 16(9):1887–1915, 2004 (cit. on p. 18).

[48] Gunning David. «Explainable artificial intelligence (xai)». In: *Defense Advanced Research Projects Agency (DARPA), nd Web 2.2 (2017)* (cit. on p. 18).

[49] Lundberg Scott and Su-In Lee. «A unified approach to interpreting model predictions». In: arXiv preprint arXiv:1705.07874 (2017) (cit. on p. 18).

[50] Sergiu Hart. «Shapley Values». In: *The New Palgrave: A Dictionary of Economics, edited by John Eatwell, Murray Milgate, and Peter Newman*. London: Macmillan, 1987 (cit. on p. 18).

[51] «Fairness (machine learning) Equità (apprendimento automatico)». In: URL: `https://it.qaz.wiki/wiki/Fairness_(machine_learning)#Relationships_between_definitions` (cit. on pp. 20, 21).

[52] Paolo Eccher. «Strumento per la valutazione di modelli di Machine Learning». In: Tesi di Laurea, Università degli Studi di Padova, 2018. Relatore: Lamberto Ballan (cit. on p. 20).

[53] Luigi Saetta. «Metriche di prestazione di un modello». In: URL: `https://luigisaetta.it/index.php/machine-learning/22-metriche-di-prestazione-di-un-modello` (cit. on p. 20).

[54] Sam Corbett-Davies, Emma Pierson, Avi Feller, and Sharad Goel. «A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear». In: *The Washington Post. Retrieved January 1, 2018.* 2016 (cit. on pp. 23, 35).

[55] Ross E. *Intelligent User Interfaces: Survey and Research Directions*. Technical Report CSTR-00-004 2000. University of Bristol, 2000 (cit. on p. 28).

[56] Fiebrink R., PR Cook, and D Trueman. «UHuman model evaluation in interactive supervised learning». In: *Proceedings of the ACM Conference on Human Factors in Computing Systems (2011),147-156* (cit. on p. 28).

[57] Sundar S and Nass C. «Conceptualizing sources in online news». In: *Journal of Communication 51 (1): 52–72*. Queue 11 (3): 1–19. Sweeney L (2013) Discrimination in online ad delivery, 2001 (cit. on p. 29).

[58] Min Kyung Lee. «Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management». In: (2018). Big Data Society 5, 1 (2018), 2053951718756684 (cit. on p. 29).

[59] Malcolm Ware, Eibe Frank, Geoffrey Holmes, Mark Hall, and Ian H. Witten. «Interactive machine learning: letting users build classifiers». In: *International Journal of Human-Computer Studies 55, 3 (2001), 281–292*. 2001. URL: https://doi.org/10.1006/ijhc.2001.0499 (cit. on p. 29).

[60] Jerry Alan Fails and Jr Dan R. Olsen. «Interactive Machine Learning». In: *Proceedings of the 8th International Conference on Intelligent User Interfaces (IUI '03)*. ACM, New York, NY, USA, 2003, pp. 39–45. URL: https://doi.org/10.1145/604045.604056 (cit. on p. 29).

[61] JSaleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. «Effective end-user interaction with machine learning». In: *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*. AAAI Press, 2011a, pp. 1529–1532 (cit. on p. 30).

[62] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. «Interacting meaningfully with machine learning systems: Three experiments». In: *International Journal of Human-Computer Studies 67, 8 (2009), 639 - 662*. 2009. URL: https://doi.org/10.1016/j.ijhcs.2009.03.004 (cit. on p. 30).

[63] JJ Dudley and PO Kristensson. «A Review of User Interface Design for Interactive Machine Learning». In: (2018). ACM Trans. Interact. Intell. Syst. (TiiS) 2018, 8, 8. [CrossRef] (cit. on p. 30).

[64] JR Venable. «The Role of Theory and Theorising in Design Science Research». In: *irst International Conference on Design Science Research in Information Systems and Technology*. Claremont, Calif., 2006, pp. 1–18 (cit. on p. 30).

[65] AR Hevner, ST March, J. Park, and S Ram. «Design Science in Information Systems Research». In: *First International Conference on Design Science Research in Information Systems and Technology*. MIS Quarterly 28 (1) (2004), 75-105. (cit. on p. 30).

[66] GM Olson and JS Olson. «Human-Computer Interaction: Psychological Aspects of the Human Use of Computing». In: *Annual Review of Psychology 54, 491–516, (2003), 491–516* (cit. on p. 30).

[67] Adikari, Sisira Mcdonald, Craig Collings, and Penny. «A design science approach to an HCI research project». In: *ACM International Conference Proceeding Series. 206, 429-432. 10.1145 / 1228175.1228265*. 2006 (cit. on p. 30).

[68] Kayur Patel, James Fogarty, James A. Landay, and Beverly Harrison. «Investigating statistical machine learning as a tool for software development». In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08). 667–676*. 2008. URL: http://dx.doi.org/10.1145/1357054.1357160 (cit. on p. 30).

[69] Steven P. Dow, Alana Glassco, Jonathan Kass, Melissa Schwarz, Daniel L. Schwartz, and Scott R. Klemmer. «Parallel prototyping leads to better design results, more divergence, and increased self-efficacy». In: *ACM Transactions on Computer-Human Interaction 17, 4 (Dec. 2010), 1–24.* 2010. URL: `http://dx.doi.org/10.1145/1879831.1879836` (cit. on p. 31).

[70] Bill Buxton. «Sketching user experiences: Getting the design right and the right design». In: (2007). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. (cit. on p. 31).

[71] Marco Gillies et al. «Human-centered machine learning». In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHIâ € ™ 16): 3558-3565.* 2016 (cit. on p. 31).

[72] P Langley. «Machine learning for adaptive user interfaces». In: *Proceedings of the 21st German Annual Conference on Artificial Intelligence, 53-62. Freiburg, Germany: Springer.* 1997 (cit. on p. 32).

[73] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann. «Mastering the information age - solving problems with visual analytics». In: (2010) (cit. on p. 32).

[74] M. Sedlmair, M. Meyer, and T. Munzner. «Design study methodology: Reflections from the trenches and the stacks». In: (2012). IEEE Trans. Vis. Comput. Graph., 18 (12): 2431–2440 (cit. on p. 32).

[75] D. Sacha, M. Sedlmair, L. Zhang, JA Lee, D. Weiskopf, SC North, and DA Keim. «Human-Centered Machine Learning Through Interactive Visualization: Review and Open Challenges, 665». In: *24th European Symposium on Artificial Neural Networks, ESANN, 2016, pp. 641–646* (cit. on p. 32).

[76] Yunkyung Kim and Bilge Mutlu. «How social distance shapes human-robot interaction». In: *International Journal of Human-Computer Studies 72: 783-795.* 2014 (cit. on p. 33).

[77] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, Siddhartha Srinivasa, and Paul E Rybski. «Gracefully mitigating breakdowns in robotic services». In: *5th ACM / IEEE International Conference on Human-Robot Interaction (HRI): 203-210.* 2010 (cit. on p. 33).

[78] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. «UX Design Innovation: Challenges for Working with Machine Learning as a Design Material». In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. ACM, 278–288. 10.1.* 2017 (cit. on p. 33).

[79] «Jupyter Notebook». In: URL: `https://jupyter.org/` (cit. on p. 34).

[80] Will Koehrsen. «Interactive Controls in Jupyter Notebooks». In: *In Toward data science*. 2019. URL: https://towardsdatascience.com/interactive-controls-for-jupyter-notebooks-f5c94829aee6 (cit. on p. 34).

[81] Chakri Cherukuri. «Jupyter Notebooks: Interactive Visualization Approaches». In: *InfoQ*. 2019. URL: https://www.infoq.com/presentations/ml-models-jupyter/ (cit. on p. 34).

[82] Jeffrey M. Perkel. «Why Jupyter is data scientists' computational notebook of choice». In: 2018. URL: https://www-nature-com.ezproxy.biblio.polito.it/articles/d41586-018-07196-1 (cit. on p. 34).

[83] Morphocode. «Interactive notebooks for data analysis and visualization». In: URL: https://morphocode.com/interactive-notebooks-data-analysis-visualization/ (cit. on p. 34).

[84] «ipywidgets». In: URL: https://ipywidgets.readthedocs.io/en/stable/ (cit. on p. 34).

[85] «pandas.DataFrame». In: URL: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html (cit. on p. 35).

[86] «Plotly». In: URL: https://plotly.com/ (cit. on p. 39).

[87] «ISO 9241-11:2018». In: URL: https://www.iso.org/standard/63500.html#:~:text=ISO%5C%209241%5C%2D11%5C%3A2018%5C%20provides,services%5C%20(including%5C%20technical%5C%20and%5C%20personal (cit. on p. 46).

[88] Norman, Donald A., Draper, and Stephen W. «User Centered System Design; New Perspectives on Human-Computer Interaction». In: *L. Erlbaum Associates Inc.* isbn:0898597811. USA, 1986 (cit. on p. 46).

[89] De Angeli Antonella. «Misure di qualità: dall'usabilità all'esperienza utente». In: 2008 (cit. on p. 46).

[90] «ReactJS». In: URL: https://it.reactjs.org/ (cit. on p. 46).

[91] Aggarwal S. «Modern web-development using reactjs». In: *International Journal of Recent Research Aspects*. Vol. 5(1). 2018, pp. 2349–7688 (cit. on p. 46).

[92] «DOM». In: URL: https://www.html.it/guide/guida-dom/ (cit. on p. 46).

[93] «Flask». In: URL: https://flask.palletsprojects.com/en/1.1.x/ (cit. on p. 47).

[94] «Jinja Template Engine». In: URL: https://jinja.palletsprojects.com/en/2.11.x/ (cit. on p. 47).

[95] «Werkzeug WSGI Toolkit». In: URL: https://werkzeug.palletsprojects.com/en/1.0.x/ (cit. on p. 47).

[96] Taneja, Sheetal, and Pratibha R. Gupta. «Python as a tool for web server application development». In: *Int. J. Information, Commun. Comput. Technol 2.1.* 2347-7202. 2014 (cit. on p. 48).

[97] Aslam, Fankar Armash, Hawa Nabeel Mohammed, and P. Lokhande. «Efficient way of web development using python and flask». In: *International Journal of Advanced Research in Computer Science 6.2.* 54-57. 2015 (cit. on p. 48).

[98] Ronacher Armin. «Flask: web development, one drop at a time». In: Retrieved May 1 (2015). 2015 (cit. on p. 48).

[99] DuPlain Ron. «Instant Flask Web Development». In: 1st ed. Olton: Packt, Limited, 2013. Web (cit. on p. 48).

[100] Copperwaite Matt and Charles Leifer. «Learning Flask Framework». In: Packt Publishing Ltd, 2015 (cit. on p. 48).

[101] Mario Mancini. «Bauhaus: alle origini dell'ossessione di Steve Jobs per il design». In: 2019. URL: https://marioxmancini.medium.com/bauhaus-alle-origini-dellossessione-di-steve-jobs-per-il-design-de14832796e7 (cit. on p. 48).

[102] WEMEDIA. «La filosofia di Steve Jobs». In: URL: https://www.wemedia.it/la-filosofia-di-steve-jobs-44.html (cit. on p. 48).

[103] R Oppermann. «User-interface design». In: *In Handbook on information technologies for education and training.* Springer, Berlino, Heidelberg. 2002, pp. 233–248 (cit. on p. 48).

[104] A Marcus. «Graphic Design for Electronic Documents and User Interfaces». In: ACM Press. 1992 (cit. on p. 50).

[105] T Boyle. «User-interface design». In: *In Handbook on information technologies for education and training.* London et al.: Prentice Hall. 1996 (cit. on p. 51).

[106] Saleiro Pedro, Kuester Benedict, Stevens Abby, Anisfeld Ari, Hinkson Loren, London Jesse, and Ghani Rayid. «Aequitas: A Bias and Fairness Audit Toolkit». In: *arXiv preprint arXiv:1811.05577* (2018) (cit. on pp. 94–97).

[107] «COMPAS Analysis using Aequitas». In: (). URL: https://dssg.github.io/aequitas/examples/compas_demo.html#disparity_calc (cit. on pp. 96, 97).

[108] Yeounoh Chung. «Slice Finder». In: (). URL: https://github.com/yeounoh/slicefinder (cit. on pp. 99, 100).