POLITECNICO DI TORINO

Master's Degree in Computer Engineering

Master's Degree Thesis

Development of a visual data annotation system based on Bayesian inference

Supervisors

Prof. ANTONIO VETRO'

Prof. JUAN CARLOS DE MARTIN

Candidate

Behnam LOTFI

April 2023

Summary

Machine learning has made significant advancements in recent years and has become a crucial tool in various domains such as credit scoring, crime prediction, and college admission. However, its results can be flawed if the underlying data is partial or incomplete. To evaluate the fairness of data, various statistical measures such as mean, median, mode, range, variance, standard deviation, skewness, kurtosis, and outliers are used to determine the distribution of values in a dataset. Nevertheless, fairness assessment is a complex task that may necessitate considering factors outside the dataset, such as the data's source and context of use.

A strategy has been proposed to raise awareness of the impact of sampling practices and to emphasize the role of data structure in the risk of producing discriminatory results. This method has been tested and has been shown to assess the danger of racial discrimination in widely used datasets. To mitigate this risk, a method of data annotation using Bayesian statistical inference has been proposed.

Bayesian statistical inference is a method of statistical inference that uses Bayes' theorem to update the probability of a hypothesis when new data becomes available. In Bayesian inference, probabilities are considered as degrees of belief and are updated using the Bayesian updating procedure. The proposed method emphasizes the significance of awareness about the sampling practices used to create the training set and the dependency of the probability of success or failure for minority groups on the structure of the available data. A system has been developed to implement this method and has been used to analyze three commonly used datasets in the machine learning community to assess the risk of racial discrimination by the system.The metrics being discussed are the essential elements for evaluating the quality of a software system. These metrics include Functional Suitability, Reliability, Performance Efficiency, Usability, Security, Compatibility, Maintainability, and Portability. These metrics serve as a benchmark for measuring the effectiveness of a system. They play a critical role in determining the overall performance and reliability of the system and ensure that it is well-equipped to perform its intended functions while also being secure and easy to use. Additionally, these metrics also assess the software's compatibility with other systems, its ability to be maintained and updated, and its portability, making it adaptable to different environments.

Acknowledgements

I would like to convey my sincere and heartfelt appreciation and gratitude to my supervisors, Professor Antonio Vetro' and Juan Carlos De Martin, for allowing me to do research and work under their supervision.

My appreciation and affection go to my parents for their support and love throughout my entire life.

Finally, the I want to express my thanks to my friends from the Polytechnic University of Turin for helping me complete my studies and thesis. I am expressing gratitude and appreciation to everybody who has had a significant impact on my life and academic journey.

Table of Contents

List of Tables VIII

List of Figures IX

1 Introduction 1

1.1 Detect discriminatory risk through data annotation 2

1.2 Bayesian statistical inference 2

1.3 Thesis objectives 3

1.4 Utilized datasets 3

1.4.1 Compas dataset 3

1.4.2 Drug consumption dataset 4

1.4.3 Adult dataset 4

1.5 Thesis structure 4

2 Literature Review 6

3 Technological architecture and design overview 9

3.1 Aims and objectives 10

3.1.1 Functional Suitability 10

3.1.2 Reliability 11

3.1.3 Performance Efficiency 11

3.1.4 Usability 12

3.1.5 Learn-ability 12

3.1.6 Operability 12

3.1.7 User Interface Aesthetics 12

3.1.8 Accessibility 13

3.1.9 Security 13

3.1.10 Maintainability 13

3.1.11 Compatibility 13

3.1.12 Portability 14

3.2 Tools and Methodologies 14

3.3 Expected outcomes 14
3.4 Functional Requirements 15
3.5 Non-Functional Requirements 15
3.5.1 Functional Requirements Use Cases 16
3.6 Diagrams 18

4 Outcomes of the application's sample datasets 21
4.1 Application Environment 21
4.2 Current Configurations 27
4.2.1 Dependence evaluation 27
4.2.2 Diverseness measurement 28
4.2.3 Inclusiveness assessment 29
4.2.4 Training Likelihood evaluation 29
4.3 Results 30
4.3.1 Results of adults dataset 31
4.3.2 Results of compass dataset 34
4.3.3 Results of drug dataset 38

5 Conclusions and potential future development 43

Bibliography 47

List of Tables

3.1 Functional Requirements 15
3.2 Non-Functional Requirements 16
3.3 Use case 1, UC1 - User upload the file of related dataset from user . 16
3.4 Use case 2, UC2 - System extracts columns from the file of dataset 16
3.5 Use case 3: User should select the required columns for evaluation . 17
3.6 Use case 4, Application executes the selected evaluation and get the results by charts and tables 17
3.7 Use case 5, User can configure application for all the application . . 18
3.8 Use case 6, Application will report as table and graph 18

List of Figures

3.1 High level design 15
3.2 Context Diagram 18
3.3 Use case diagram 19
3.4 Deployment Diagram 19
3.5 Class diagram 20
3.6 Sequence diagram 20

4.1 Main page of application 22
4.2 Upload page of application initial view 22
4.3 Upload page of application file selected 23
4.4 Upload page of application file uploaded 23
4.5 Select uploaded file to use for evaluations or delete file 24
4.6 Select columns or features of dataset for evaluations 24
4.7 Possibility of selection of module for calculation 25
4.8 Available modules for calculations 25
4.9 Select desirable values for calculations and calculate 26
4.10 Running by get results button after loading data 26
4.11 Depenence sample results after running 27
4.12 Adults income diverseness 31
4.13 Adults income diverseness chart 31
4.14 Adults race income dependence 32
4.15 Adults race income dependence Chart 32
4.16 Adults race income inclusiveness 32
4.17 Adults race income inclusiveness chart 33
4.18 Adults race income likelihood 33
4.19 Adults race income likelihood chart 34
4.20 COMPAS race diverseness 34
4.21 COMPAS race diverseness chart 35
4.22 COMPAS race score dependence 35
4.23 COMPAS race score dependence Chart 36
4.24 COMPAS race score inclusiveness 36

4.25 COMPAS race score inclusiveness chart 37
4.26 COMPAS race score likelihood 37
4.27 COMPAS race score likelihood Chart 38
4.28 Drugs CL2 Diversness 38
4.29 Drugs CL2 CL0 Diversness Chart 39
4.30 Drugs CL2 CL0 dependence 39
4.31 Drugs CL2 CL0 dependence chart 40
4.32 Drugs CL2 CL0 inclusiveness 40
4.33 Drugs CL2 CL0 inclusiveness chart 41
4.34 Drugs CL2 CL0 likelihood 41
4.35 Drugs CL2 CL0 likelihood chart 42

Chapter 1

Introduction

Machine learning (ML) has progressed dramatically in recent decades and become a useful technique in a variety of applications, including credit scoring [1], crime prediction [2] and college admission [3][4]. The rapid advancement in machine learning research, due to the increase in computational power and data availability, has led to the widespread use of these systems in decision making. However, it is well known that these systems can produce problematic results if they are based on partial or incomplete data.

A variety of statistical factors can be evaluated when evaluating data to figure out if it is fairly distributed statistically. Mean, median, and mode are helpful measures for comprehending the central tendency of a collection and the distribution of values surrounding the central value. Range, variance, and standard deviation are dispersion measurements that can be used to determine the value distribution within a set. The skewness of a dataset is a measure of its asymmetry. Positive skewness indicates a longer tail in the positive direction, and negative skewness indicates a longer tail in the negative way. Kurtosis is a measurement of a dataset's peakiness. The distribution of data sets with a high kurtosis has more peaks than those with a low kurtosis. Outliers are values in a dataset that differ significantly from the rest of the data[5]. It is crucial to identify and appreciate the significance of outliers while analyzing a dataset, since their existence can have a significant impact on the measures of central tendency and dispersion. It is vital, while analyzing fairness, to examine the distribution of values across different categories, such as demographic groupings, to establish whether the data is spread equitably throughout these groups. For instance, the mean or median of a group can be compared to the mean or median of the entire dataset to assess if the group is fairly represented. These statistical characteristics can provide useful insights into the distribution of values within a dataset and can be used to uncover potential concerns around fairness. However, it is essential to keep in mind that assessing data for fairness is a challenging task that may need consideration of factors outside the distribution of values in the dataset, such as the data's sources and context of use. Although recent studies have highlighted the ethical and transparency concerns related to data collection and recording in these systems and numerous criteria, some of which are conflicting, have been published in recent years to evaluate the fairness of classification and regression models.

1.1 Detect discriminatory risk through data annotation

Discriminatory risk in data annotation refers to the possibility that a model trained on annotated data will produce predictions that discriminate or unfairly treat particular groups of individuals on the basis of their demographic features, such as race, gender, or age. This can arise when the annotated data used to train the model reflects biases in the data-generation process or in the annotators, resulting in predictions that perpetuate or magnify these biases.

For instance, if a model is trained on annotated data in which the majority of examples of a particular occupation are of one gender, the model may learn to link that occupation with that gender, even though the actual relationship between occupation and gender is not intrinsically discriminatory. This might lead to inaccurate predictions when the model is applied in the real world, such as recommending specific careers to individuals based on their gender[6].

In the realm of machine learning and artificial intelligence, discriminatory risk in data annotation is an increasing concern, and researchers and practitioners are actively developing strategies to mitigate these risks and ensure that machine learning models are fair and equitable. This may involve utilizing diverse and representative annotated data, employing fairness-aware algorithms, and evaluating the performance of models on various demographic groups in order to identify and resolve any potential biases.

1.2 Bayesian statistical inference

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to revise the probability of a hypothesis when additional evidence or data becomes available[7]. It is predicated on the notion that probabilities can be used to express our uncertainty regarding the events or hypotheses in question, and that this uncertainty can be updated when new information is gathered.

In Bayesian inference, probabilities are considered as degrees of belief and are updated using the Bayesian updating procedure. This method of updating entails calculating the posterior probability of a hypothesis given the available data and prior probability, which is our initial belief about the hypothesis before evaluating the new evidence. In subsequent updates, the posterior probability is utilized as the new prior probability, integrating any new data that becomes available.

Bayesian inference provides a flexible framework for combining previous knowledge and modeling complex interactions between variables, and has been utilized in numerous applications, such as machine learning, signal processing, and decision making.

1.3 Thesis objectives

A method of data annotation utilizing Bayesian statistical inference has been developed to address the potential for unequal treatment of specific groups of individuals due to discriminating risk. This strategy aims to uncover any potential biased results within a given dataset and to raise awareness of the sampling techniques used to create the training set. It also underscores the fact that the structure of available data may significantly lead to the success or loss of minority groups. Based on this methodology, a system has been constructed, and three widely used datasets in the machine learning community have been investigated and appraised for the danger of racial discrimination. Datasets the compas [8], the drug consumption [9], [10] and the adult [9] are examples of widely used datasets in machine learning.

1.4 Utilized datasets

1.4.1 Compas dataset

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)[11] is a popular commercial algorithm used by courts and parole officers to score the likelihood of recidivism of criminal defendants (recidivism). A 2-year follow-up research demonstrates that the algorithm is skewed in favor of white defendants and against black inmates (i.e who actually committed crimes or violent crimes after 2 years). As assessed by precision/sensitivity, the error pattern is remarkable. This dataset contains variables used by the COMPAS algorithm in scoring defendants, along with their outcomes within 2 years of the decision, for over 10,000 criminal defendants in Broward County, Florida and information about individuals, including demographic information, and can be used to determine if an algorithm is biased towards or against certain groups based on race, gender, age, etc.

1.4.2 Drug consumption dataset

Datasets on drug consumption are frequently used for fairness analysis since they can provide insight into drug usage trends and associated effects across demographic groups. This data can assist academics and policymakers in identifying potential causes of bias and discrimination in the criminal justice system, the healthcare system, and other areas where drug use may influence individuals and communities. By studying the distribution of drug use and related consequences across different demographic groups, researchers can uncover inequities and attempt to eliminate them through the implementation of interventions and policies supported by scientific evidence. This can aid in promoting fairness and equity in the treatment of drug users, as well as reducing the harm associated with drug use and dependency.

1.4.3 Adult dataset

Adult datasets are frequently used for fairness research because they provide information on demographic features, employment status, income, level of education, and other critical aspects that can influence an individual's access to opportunities and life outcomes. By evaluating these data, researchers can determine whether inequities or biases exist in various systems, such as the job market, housing, education, and healthcare.

By assessing the representation of different demographic groups in different job categories and the drivers of work performance, for instance, an adult dataset can be used to evaluate the fairness of a recruiting process. If a certain group is underrepresented in higher-paying positions or has a lower likelihood of promotion, this may suggest bias in the hiring process.

Likewise, adult datasets can be utilized to evaluate the fairness of lending practices by analyzing the distribution of loans and credit across various demographic groups. If a certain group is less likely to be approved for loans or is charged higher interest rates, this may be an indication of discrimination in the lending process.

Using this dataset researchers and policymakers can uncover the roots of inequities and biases and implement evidence-based initiatives to increase justice and equality in many systems by analyzing adult datasets for fairness.

1.5 Thesis structure

The thesis outlines its topic, objective, and informational flow of the thesis in the beginning. This section provides an overview of the main theme and purpose of the research.

In chapter two, a literature review is provided. This section reviews relevant literature and studies that have been previously conducted on the topic. This review helps to provide context and background information on the subject and provides a foundation for the research being conducted.

Chapter three focuses on the structure and design of the system that is being studied. This section provides an in-depth analysis of the architecture and methodology used to develop the system. The structure and design of the system are critical in understanding its functionality and overall performance.

Chapter four presents the results of the system for commonly used sample datasets in data science. This section provides insight into the performance and accuracy of the system and demonstrates its capabilities. The results of the system are important in evaluating its effectiveness and validity.

Finally, possible ways to extend the functionality of the application and future research directions are proposed at the end of the thesis. This section provides recommendations for future improvements to the system and potential areas for further research.

Chapter 2

Literature Review

Several publications are now examining the dangers of discrimination in machine learning and referencing relevant research. These citations are likely to be works that investigate bias in AI and its potential social consequences. These interrelated works address the root causes of inequality in machine learning, its potential outcome on individuals and communities, and potential remedies to this issue.

In the fall of 2020, Twitter users raised complaints that Twitter's automated image cropping mechanism favored light-skinned folks over dark-skinned individuals and favored cropping woman's bodies rather than their heads. Twitter uses machine learning to crop images, where crops are centered around the part predicted to be the most salient. To address these issues, A comprehensive analysis of formalized group fairness indicators was conducted. Then systematic differences in cropping and identify significant causes, including the fact that cropping based on the single most important point might increase the differences due to a phenomenon discovered that named as argmax bias. Nonetheless, it is shown that established fairness standards and quantitative analysis alone are insufficient to capture the possibility of representational harm in automatic cropping. it was recommended removing saliency-based cropping in favor of a technique that preserves user agency more effectively. that evaluation inspires a combination of quantitative and qualitative methodologies, including human-centered design, for the development of a new solution that adequately addresses concerns about representational harm [12] [13].

There is additional research on the Facebook problem[14]. The article discusses the issue of racially discriminatory advertising on Facebook and its resolution. Racially discriminatory advertising refers to advertisements that use racial profiling or target specific races, which can lead to unequal treatment and is illegal in many nations. The article may examine several remedies to this problem, such as implementing stricter norms and regulations, utilizing technology to monitor and detect discriminatory advertisements, or educating advertisers on diversity and inclusion. The article also provide insights and solutions for more effectively addressing this issue using auditing algorithms for fairness[15].The study describes a methodology for measuring the delivery of Facebook ads along racial lines. The authors aim to run groups of ads and vary a particular feature to see how changing the feature affects the set of users that Facebook delivers the ad to. To control which users are in the target audience, the authors use random PII-based custom audiences where they randomly select US Facebook users to be included. They generate custom audiences by randomly generating 20 lists of 1 million distinct US phone numbers.

In the research conducted by Lui et al.[16][17], Bayesian analysis is used to establish the link between default probability and time. Using mixture models to represent the unnamed subgroup, their efforts to find default time were successful. A combination of hierarchical and non-hierarchical models were utilized to demonstrate the effectiveness of the system for banks in better understanding individual behavior. Their results are encouraging for providing lenders with timely capital risk management indicators. Their solution incorporated Markov Chain Monte Carlo without variable selection, time-based variables such as the prime (variable) interest rate, and a Bayesian-based system that was efficient in default timing.

The previous articles [18] presented their findings independently. Nonetheless, the present application has been intended to give a more universal process for calculating and annotating data, as well as the option to add more computations to the process. In addition, the article "Detecting Discrimination Risk in Automated Decision-Making Systems using Balance Measures on Input Data"[19] is utilized as input for this application. This article explains several indexing techniques that can be included into the existing software via configuration files.

While some systems, such as "A framework for clinical data integration and annotation for decision support,"[20] are designed especially for particular data sets, the current application offers a broader field of use. Adaptable input configurations permit this program to give valuable information and calculations for a wider range of data sets.

There is an article for comparing Bayesian model which [21] is referring to other studies or research papers that are exploring the use of Bayesian Models of Annotation to address the issue of discriminatory risk. The Bayesian Models of Annotation is a statistical method used to analyze and interpret data and is being considered as a solution to the problem of discriminatory risk. In these studies, the Bayesian Models of Annotation is being compared to other approaches or methods to determine its effectiveness in reducing or mitigating discriminatory risk.

Importantly, previous papers lacked the thorough and consistent methodology supplied by the current application. The ability to compute and annotate data in a uniform manner, while also having the flexibility to add extra calculations, sets this application apart from the others. In addition, the present program may combine data from other papers and can be customized to function with a wider variety of data sets, making it a powerful data analysis tool.

In conclusion, the new program offers a superior alternative to prior ways because it provides a consistent mechanism for calculating and annotating data, as well as the possibility to add additional computations and include data from other sources. In addition, its ability to adapt to varied data sets make it a powerful tool for data analysis and an asset in a variety of sectors.

Chapter 3

Technological architecture and design overview

3.1 Aims and objectives

The overall goal of this thesis is to construct and demonstrate a system that can be used to deliver an annotation data system, and the following goals have been set to achieve:

1. The thesis work's primary aim would be to resemble the paper's [18] visualizations and metrics using the identical data sets. It would be in the form of a web application to get content considering. This proposal would verify an overview of eight product quality characteristics and 31 sub-characteristics in accordance with ISO/IEC 25010:2011 [22]:

    (a) Functional Suitability

    (b) Reliability

    (c) Performance Efficiency

    (d) Usability

    (e) Security

    (f) Compatibility

    (g) Maintainability

    (h) Portability

2. The thesis' second focus is on providing interactivity to the visualizations, based on a collection of user stories provided by advisors.

3. As a third stage, providing some degree of generalizability, such as the ability to add additional datasets from the user and select columns for analysis or other to be agreed upon.

3.1.1 Functional Suitability

Functional Suitability: This refers to a product's or system's ability to offer functions that fulfill both stated and implicit requirements. Both R and Python, have the ability to add additional features via plugins and libraries. Furthermore, the R language benefits from the usage of more particular statistical libraries, whereas Python tends to make use of a wide range of programming libraries.

Functional Completeness: The set of functions that covers all the defined tasks and user objectives is referred to as functional completeness. Because both environments include numerous third-party libraries, it's difficult to evaluate their functional completeness because they have distinct aims. For R data analysis and statistics, and Python deployment and production, each of them is preferable within each section of the project.

Functional Correctness: This term refers to how successfully a product or system delivers the desired outputs with the required accuracy. There are some libraries for testing application functionality, such as PyTest in python, and both R and Python have required types in order to do computations.

Functional Appropriateness: This term refers to the ability of functions to complete specific tasks and objectives. R is more appropriate and simpler to locate essential capabilities because it is specialized to follow statistical rules and functions, while Python advantages from its reach libraries related to machine learning and statistics.

3.1.2 Reliability

This relates to how well a system, product, or component operates under specific conditions. The degree to which a system, product, or component performs specified functions under specified conditions for a specified period of time" [23] is defined in ISO/IEC 25010:2011. Maturity refers to a system's, product's, or component's capacity to satisfy its dependability requirements. Both environments include functionalities that are available even for big data to handle huge amount of data [24], such as Pyspark in Python and Sparklyr in R, and this can be utilized in both environments by help of a large number of users, so that both environments have diverse functions for different purposes.

Availability: If a system, product, or component is operational and accessible, it is said to be available. R and Python are both freely accessible and can be downloaded [25]. Both systems benefit from fault tolerance and recover-ability which are other sub-characteristics of reliability.

3.1.3 Performance Efficiency

The term "performance efficiency" denotes to how well something operates in relation to the number of resources it consumes [26]. Time Behavior refers to a product's or system's response and processing times, as well as throughput rates, while it is executing its tasks. There are certain benchmarks [26] that indicate Python is superior at calculating machine learning algorithms. Both systems, however, may be improved for greater speed by utilizing diverse libraries. The quantity and types of resources consumed by a product or system while fulfilling its functions are referred to as resource utilization. The resource use in both environments may vary based on the code implementation. The maximum limitations of a product or system parameter are referred to as capacity. Both are infinite in terms of advancement because they are open source and free, have numerous plugins, and both have strong communities for assistance.

3.1.4 Usability

The usability of a product or system relates to how well it can be utilized to fulfill certain goals effectively, efficiently, and satisfactorily. In respect of this characteristic, there is a framework called "shiny”, which is a library that provides web components for R. This framework makes it simple to create web components. Frameworks like "Dash" or "Streamlit" are alternatives to shiny in Python. The Django framework, in combination with a pure implementation [27] of the business layer, is another solution, however it is less practical in terms of usability because it is better suited to general-purpose functions. As a regular user, this application is simple to use, and as a user who can customize or alter the system, the configuration files may be used to alter the behavior.

Appropriateness Recognizability: Refers to how well you can recognize whether a product or system is appropriate for your needs.

3.1.5 Learn-ability

Refers to how easy it is to learn how to use a product or system. The learning curve is steeper for R. This implies the programmer will have to devote a significant amount of effort to studying and comprehending R code. R is also a low-level programming language, thus even basic procedures can take a long time to code [25]. Python, on the other hand, is known for its simplicity. And although there are no GUIs for it at the moment, Python’s notebooks provide great features for documentation and sharing [25].

3.1.6 Operability

This relates to whether a product or system includes attributes that make it simple to use and control[23]. Python is a multi-paradigm, high-level programming language. Python has emerged as one of the most promising languages in recent years due to its simple syntax and compatibility with a broad range of eco-systems. R is a data science language that is both wide and adaptable. R focuses on allowing users to develop algorithms and computational statistics for data analysis as a computer language [28]. User Error Protection: Refers to how well a system protects users against making errors. About programming errors both have validations for code errors but about discriminatory risk through data annotation this project is going to cover it.

3.1.7 User Interface Aesthetics

Refers to whether a user interface is pleasing. Both R and Python provide frameworks for creating web user interfaces, however Python has more flexibility when utilizing web development tools to construct unique user interfaces.

3.1.8 Accessibility

Refers to how well a product or system can be used with the widest range of characteristics and capabilities. Plugins, once again, allow both ecosystems to expand their capabilities[22].

3.1.9 Security

The security of a product or system refers to how well a product or system protects information and data from security vulnerabilities[22]. Security sub-characteristics such as confidentiality, Integrity, Non-repudiation, Accountability, Authenticity, are followed by R and python.

3.1.10 Maintainability

The capacity of a product or system to be adjusted to enhance, rectify, or adapt to changes in the environment as well as needs is referred to as maintainability. In the programming section, both environments can be diagnosed for faults, although in the case of data, some machine learning languages are complex.

Modularity refers to the ability of a system or program's components to be modified with minimum influence on the other components. To maintain modularity, R and Python employ packages and libraries. Because of the modular nature of both languages, accessible codes could be reused. Modifiability refers to a product's or system's ability to be adjusted without introducing defects or lowering its quality. Because python is used to construct generic applications such as web and desktop apps, machine learning algorithms, and so on, it delivers superior environments that can be modified much more easily than frameworks like shiny in R or Dash in python. Testability is a sub-feature of modifiability that validates functional correctness.

3.1.11 Compatibility

Compatibility refers to how well a product, system, or component can exchange information as well as perform its required functions while sharing the same hardware or software environment[22]. As the aim of implementation of current project is to divide sections of application in user interface and business and web server, therefore it can follow Co-existence and Interoperability.

3.1.12 Portability

Since R and Python are cross-platform, programs written in them may be used in a variety of contexts.

With relation to data validation, each of related features is measured using ISO/IEC 25024:2015 quality-related attributes measurements. The characteristics can be classified as "Inherent" if they are solely dependent on the data, such as completeness. They can also fall under the "System-dependent" category, such as recoverability. Efficiency, for example, might fall within both categories. In ISO/IEC 25012:2008, data imbalance is not a criterion for data quality[23].

3.2 Tools and Methodologies

The thesis tries to use different research approaches that are available in article “Detecting discriminatory risk through data annotation based on Bayesian” [18]. According to this article, methodology is based on these modules: 1. Dependence: assesses the degree of connection among the protected attributes. 2. Diverseness: provides the training diversification probability in respect to each level of the protected attribute and the target variable. 3. Inclusiveness: provides the probability that two properties are simultaneously included in the training set. 4. Training Likelihood: provides the occurrence likelihood of the protected attribute levels given the target variable levels - and vice versa - before the training set is sampled[18]. First, a study is conducted to gain fundamental understanding of data visualization and related technologies. The study embraces existing articles, books, data visualization courses [29] and web resources that were appropriate. Second, after reviewing a variety of technologies, the utilization, and characteristics of both languages in term of software quality standard are compared. Moreover, available datasets for this thesis are widely spread datasets in machine learning community, and specifically for the first and second goal of data annotation system related to this thesis, three datasets will be considered which are, the COMPAS dataset [8], the Drug Consumption dataset [9], [10] and the Adult dataset [9].

3.3 Expected outcomes

The project’s outcome after some investigations would be a web application covering goals that are introduced in this thesis. This web application is going to reach goals using tools and methodology provided in an appropriate way with the help of advisors. This project will be divided into three sections based on the characteristics described previously. The frontend is the layer that receives data from the user, whereas the web server is a Django server that can request a job from the backend or business layer.

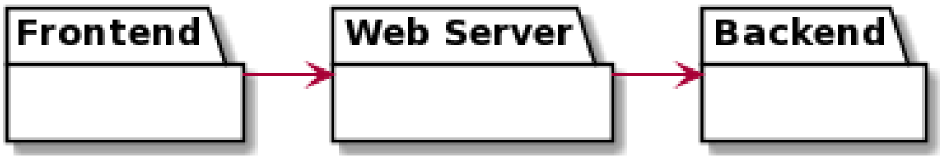


Figure 3.1: High level design

There will be three kinds of users in this project: anonymous users who can acquire random or last results for demo purposes, users who can upload datasets and request tasks from the system, and administrators who can alter configuration and algorithms as needed.

3.4 Functional Requirements

The Functional Requirements outline the expected functionality of a software system. They specify the system's behavior and the tasks it must fulfill. Typically defined in terms of inputs, outputs, and procedures, these requirements help to guarantee that the system fits the expectations of end-users and stakeholders. Functional requirements of the implemented application is listed in the table 3.1.

ID	Description
FR1	User uploads the file of data-set
FR2	System extracts columns from the file of data-set (CSV file)
FR3	User should select the required columns for evaluation
FR4	User selects the evaluations needed on data-set
FR5	Application executes the selected evaluation and get the results by charts and tables
FR6	User can configure configurations for all the application
FR7	Application reports related information regarding the data-set

Table 3.1: Functional Requirements

3.5 Non-Functional Requirements

Non-Functional Requirements (NFRs) are the quality attributes, performance standards, and design restrictions that define how well and what a software system must perform. NFRs specify the expected level of software system performance, security, scalability, usability, interoperability, and maintainability, among other non-functional characteristics. NFRs supplement functional requirements, which outline the precise activities the system must accomplish. Response time, dependability, availability, security, and data privacy are examples of non-functional requirements. Here are two non functional requirements for this application listed in the table 3.2.

ID	Description
NFR1	(Functionality) Application exports results as a file (ex. pdf)
NFR2	Anonymous user can view random evaluation and charts for demo

Table 3.2: Non-Functional Requirements

3.5.1 Functional Requirements Use Cases

In the following tables, all use cases that are implemented in the application along with their properties such as involved actors, pre and post conditions, and scenario are listed.

Actors	Involved Anonymous user
Precondition	Server is running
Post condition	File is uploaded in the server
Nominal Scenario	User uploads the file of data-set

Table 3.3: Use case 1, UC1 - User upload the file of related dataset from user

Actors	System
Precondition	File is uploaded
Post condition	Columns are ready and select-able to apply evaluations
Nominal Scenario	System extracts columns from the file of data-set (CSV file)

Table 3.4: Use case 2, UC2 - System extracts columns from the file of dataset

Actors	Involved Anonymous user
Precondition	Columns are ready and select-able to apply evaluations
Post condition	required columns are selected for evaluation
Nominal Scenario	User should select the required columns for evaluation

Table 3.5: Use case 3: User should select the required columns for evaluation

Actors Involved	Anonymous user
Precondition	Uploaded dataset, selected columns, selected evaluations
Post condition	Generated report by system
Nominal Scenario	The application runs the chosen analysis and displays the results in graphs and tables.

Table 3.6: Use case 4, Application executes the selected evaluation and get the results by charts and tables

Actors Involved	Anonymous user
Precondition	predefined evaluations configurations
Post condition	-
Nominal Scenario	If there are any predefined evaluations, the user can configure parameters for them, as well as create new evaluations if that is available.

Table 3.7: Use case 5, User can configure application for all the application

Actors Involved	Anonymous User
Precondition	Available evaluation results
Post condition	-
Nominal Scenario	Application reports related information regarding the data-set

Table 3.8: Use case 6, Application will report as table and graph

3.6 Diagrams

In the following section, context diagram, use case diagram, and class diagram are followed by sequence diagram.

A context diagram, also known as a Level 0 data flow diagram, provides an overview of a system's components, inputs, outputs, and data stores. It is used to depict the interrelationships between a system and its environment and to identify the system's boundaries. Figure 3.2 depicts the context diagram for the present application.

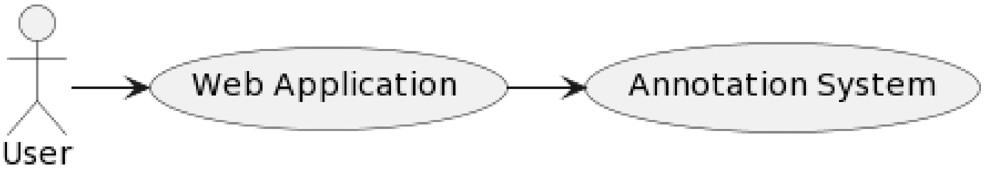


Figure 3.2: Context Diagram

Figures 3.4 and 3.5 represent the structure of the system, and figure 3.6 shows the sequence of events in the system.

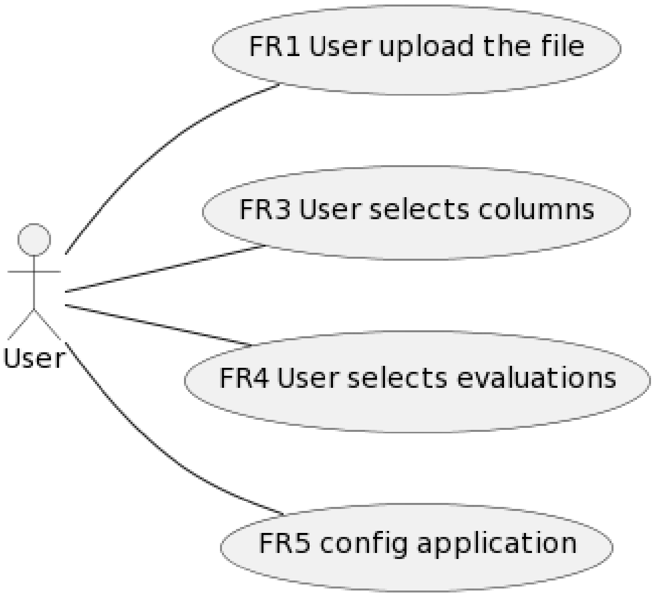


Figure 3.3: Use case diagram

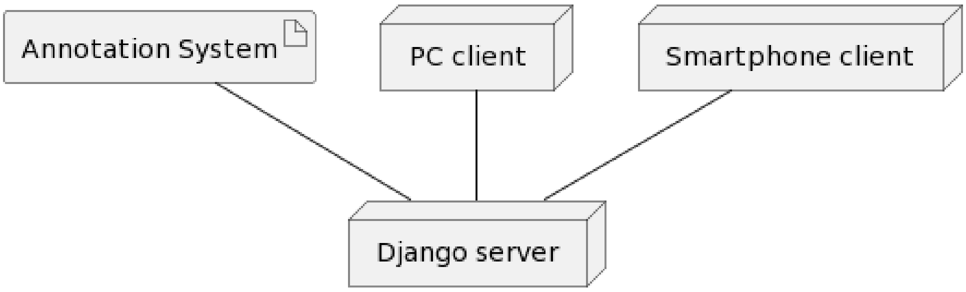


Figure 3.4: Deployment Diagram

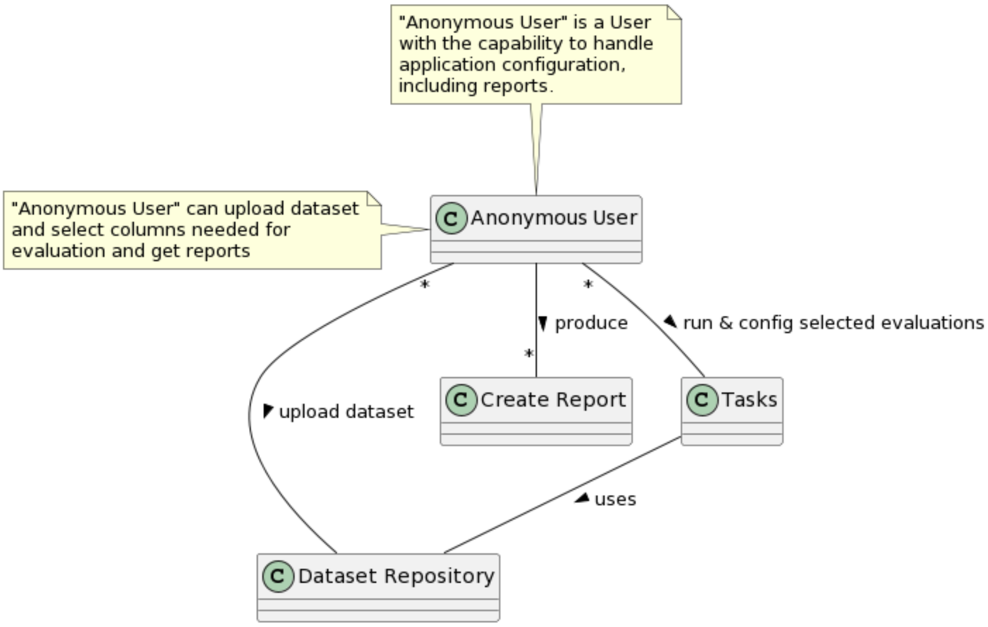

"Anonymous User" is a User with the capability to handle application configuration, including reports.
"Anonymous User" can upload dataset and select columns needed for evaluation and get reports
Anonymous User
Create Report
Tasks
Dataset Repository
upload dataset
produce
run & config selected evaluations
uses
*
*
*

Figure 3.5: Class diagram

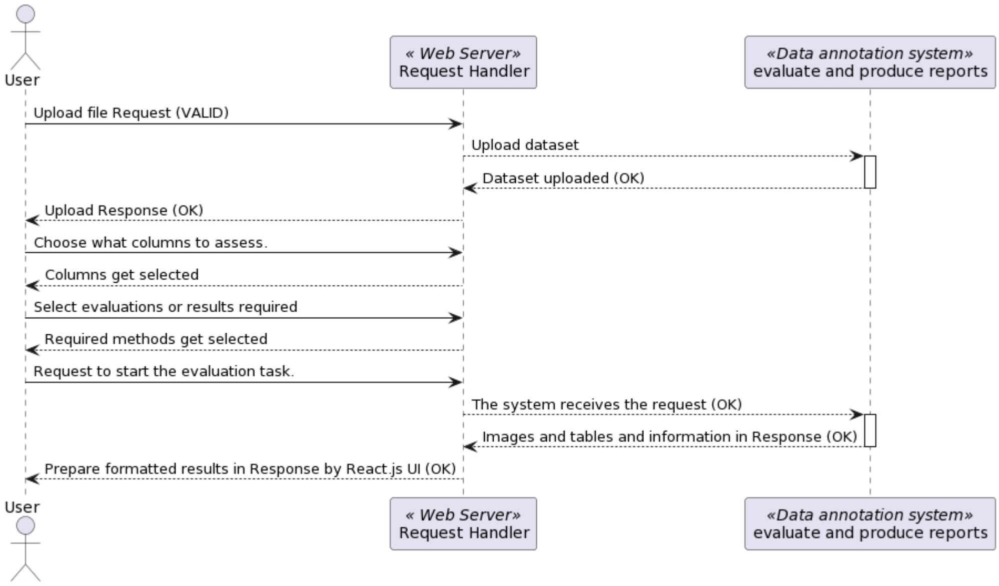

User
« Web Server»
Request Handler
«Data annotation system»
evaluate and produce reports
Upload file Request (VALID)
Upload dataset
Dataset uploaded (OK)
Upload Response (OK)
Choose what columns to assess.
Columns get selected
Select evaluations or results required
Required methods get selected
Request to start the evaluation task.
The system receives the request (OK)
Images and tables and information in Response (OK)
Prepare formatted results in Response by React.js UI (OK)

Figure 3.6: Sequence diagram

Chapter 4

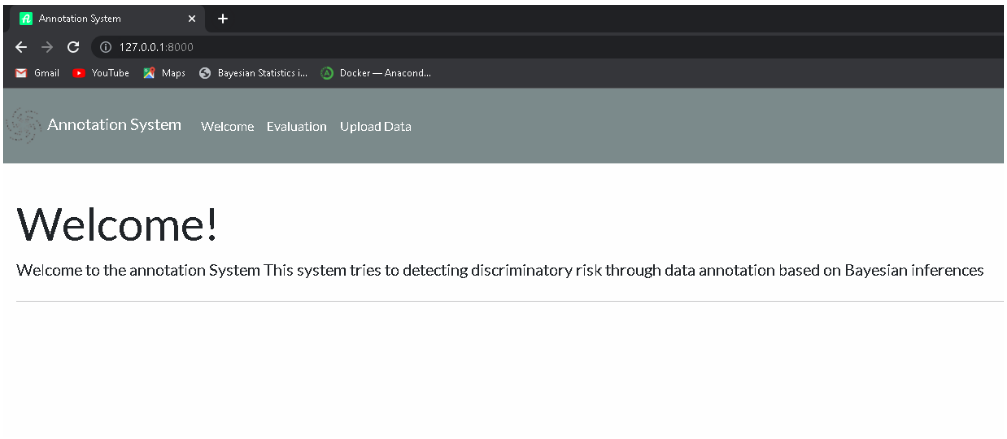
Outcomes of the application's sample datasets

4.1 Application Environment

This section will demonstrate the application environment's usage and how it can be used to various datasets and different configurations and provides results of some sample datasets.

The application is designed so that it may be easily expanded to accommodate future modules for calculations. For the applications multiple configurations for various calculations are provided. configurations consist of three individual files. One of the files has the file extension json and contains the html input elements required for each computation. For example, a user only needs a multiselect element to find the diversity of a desired request. For other calculations, such as dependence, where the source and destination values of the request must be distinguished, there are additional checkboxes with the column or feature name to specify the destination values that will be in modules or formulas. There is an additional file for storing Python formulas relevant to each module. This would be the available computations for modules such as diversity. These files have the extension.txt and contain the Python code associated with a module. The final file is named "fileList.csv" and contains the names of modules that would appear in the program, as well as the names of the json file and text file that were previously specified in order to use them in the application.

The application uses the technologies Django for the backend, React for the front end, Babbel and WebPack for packaging, and React-mui and bootstrap for ui components and styles.For ease of use, a docker container is also provided. First page you see after runing docker container on port 8000 which is provided to achive the application inside docker is the figure 4.1

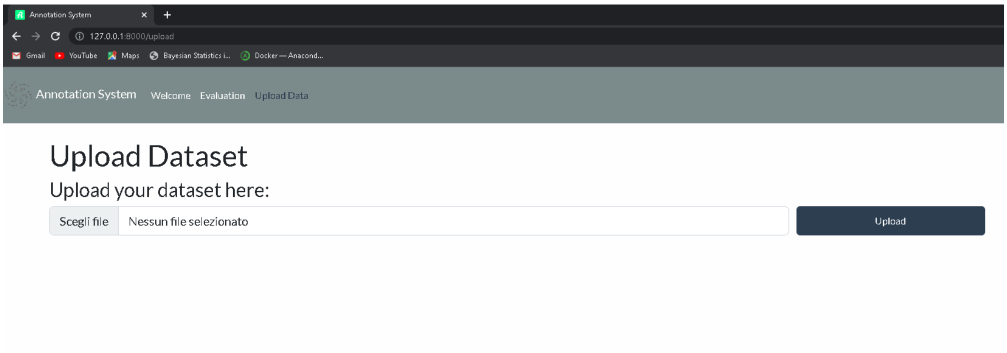


Welcome!

Welcome to the annotation System This system tries to detecting discriminatory risk through data annotation based on Bayesian inferences

Figure 4.1: Main page of application

If you are using the docker image for the first time, four datasets are supplied by default, but you can upload additional datasets in order to utilize the program to assess or compute the desired result. If a bare project is used, required datasets must be provided in order to analyze the results of uploaded datasets. Upload screen is provided in figure 4.2 and it is accessible from the menu Upload data in the application.

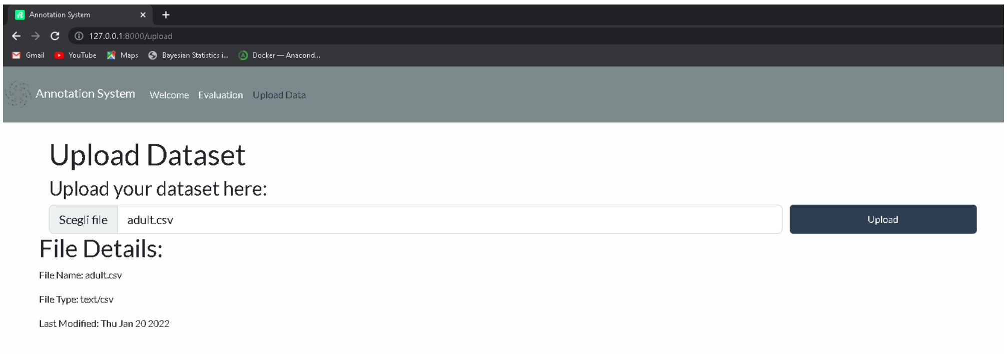


Upload Dataset

Upload your dataset here:

Figure 4.2: Upload page of application initial view

After selecting the desired file, the application will display the file's information. this is shown in the figure 4.3.



Upload Dataset

Upload your dataset here:

adult.csv

File Details:

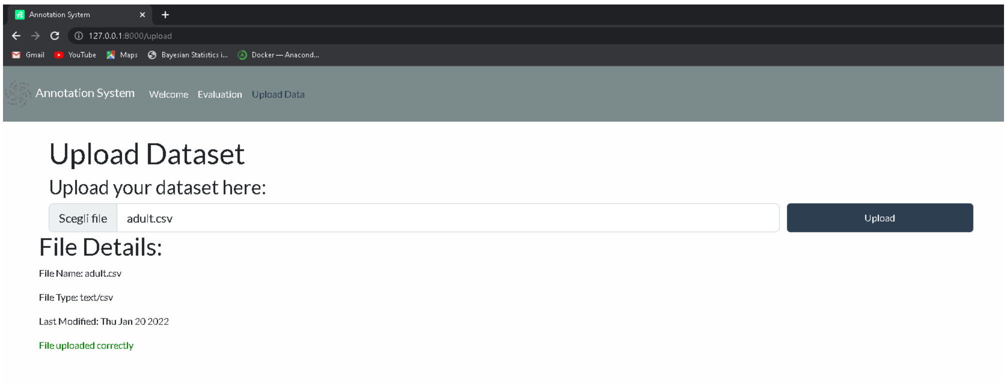
File Name: adult.csv

File Type: text/csv

Last Modified: Thu Jan 20 2022

Figure 4.3: Upload page of application file selected

By uploading the data file, which must be formatted column-wise like a csv file for the program to accept it, the application will respond with a message indicating that the file was successfully uploaded, as depicted in the figure 4.4.



Upload Dataset

Upload your dataset here:

adult.csv

File Details:

File Name: adult.csv

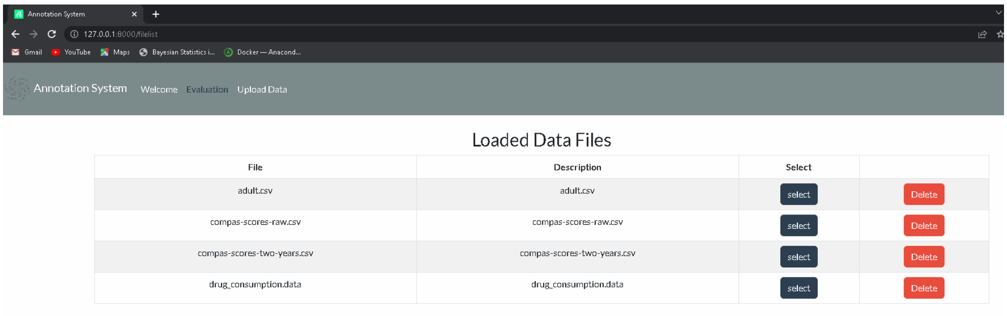
File Type: text/csv

Last Modified: Thu Jan 20 2022

File uploaded correctly

Figure 4.4: Upload page of application file uploaded

After providing the correct dataset using the upload view in the previous section, it is now feasible to apply calculations to the datasets. It is possible to choose or remove each uploaded file in this screen using select or delete buttons in front of each row like the figure 4.5.

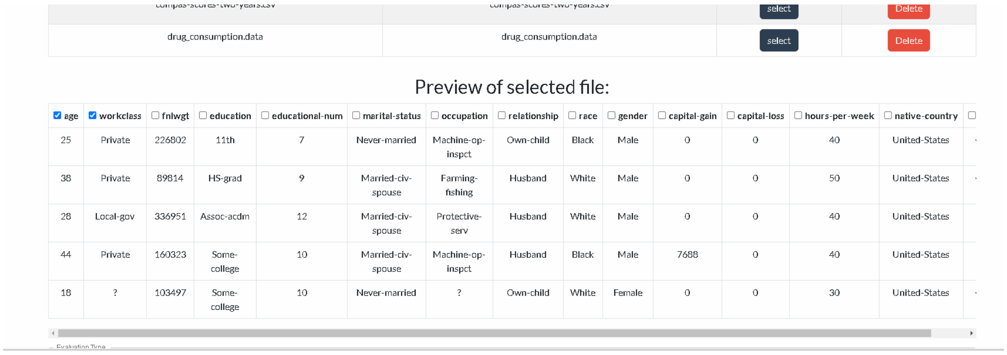


Loaded Data Files

File	Description	Select	
adult.csv	adult.csv	select	Delete
compas-scores-raw.csv	compas-scores-raw.csv	select	Delete
compas-scores-two-years.csv	compas-scores-two-years.csv	select	Delete
drug_consumption.data	drug_consumption.data	select	Delete

Figure 4.5: Select uploaded file to use for evaluations or delete file

By selecting a dataset, a preview of its initial rows will be displayed. Here, user can select columns on which computations will be performed. A sample image of the screen is provided in figure 4.6.



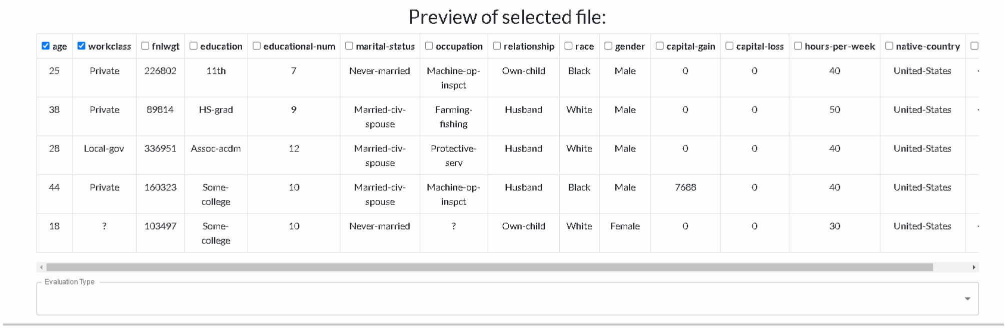
compas-scores-two-years.csv	compas-scores-two-years.csv	select	Delete
drug_consumption.data	drug_consumption.data	select	Delete

Preview of selected file:

☑ age	☑ workclass	☐ fnlwgt	☐ education	☐ educational-num	☐ marital-status	☐ occupation	☐ relationship	☐ race	☐ gender	☐ capital-gain	☐ capital-loss	☐ hours-per-week	☐ native-country	☐
25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	0	40	United-States	·
38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0	50	United-States	·
28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0	40	United-States	
44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	0	40	United-States	
18	?	103497	Some-college	10	Never-married	?	Own-child	White	Female	0	0	30	United-States	·

Figure 4.6: Select columns or features of dataset for evaluations

In the bottom portion of the figure 4.7 is a dropdown element that loads the app's available modules which is described in the beginning of the this chapter in the file called fileList.csv. Therefore, if the user adds a row to that file, a corresponding option will be added to this dropdown element.

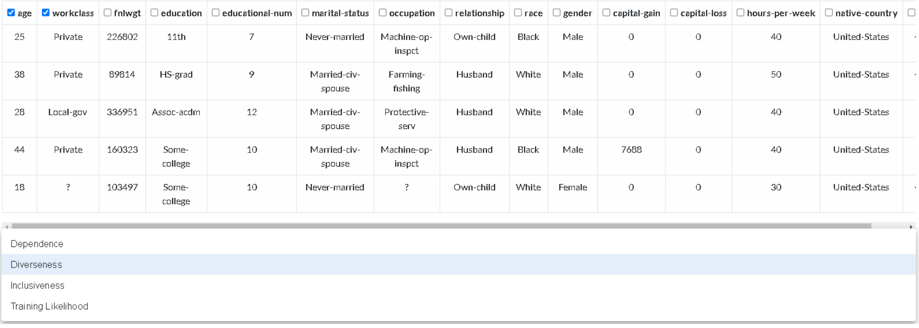


Preview of selected file:

☑ age	☑ workclass	☐ fnlwgt	☐ education	☐ educational-num	☐ marital-status	☐ occupation	☐ relationship	☐ race	☐ gender	☐ capital-gain	☐ capital-loss	☐ hours-per-week	☐ native-country	☐
25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	0	40	United-States	-
38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0	50	United-States	-
28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0	40	United-States	
44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	0	40	United-States	
18	?	103497	Some-college	10	Never-married	?	Own-child	White	Female	0	0	30	United-States	-

Figure 4.7: Possibility of selection of module for calculation

As previously explained current possible configurations are listed as it is shown in the figure 4.8.



☑ age	☑ workclass	☐ fnlwgt	☐ education	☐ educational-num	☐ marital-status	☐ occupation	☐ relationship	☐ race	☐ gender	☐ capital-gain	☐ capital-loss	☐ hours-per-week	☐ native-country	☐
25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	0	40	United-States	-
38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0	50	United-States	-
28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0	40	United-States	
44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	0	40	United-States	
18	?	103497	Some-college	10	Never-married	?	Own-child	White	Female	0	0	30	United-States	-

Figure 4.8: Available modules for calculations

After selecting the desired module for computation, as depicted in figure 4.9, the Load button will display ordered different values for each column based on the configuration for HTML components. Some include check boxes to differentiate target values for calculations, while others, such as diversity, do not, as the values serve the same purpose in the formula.

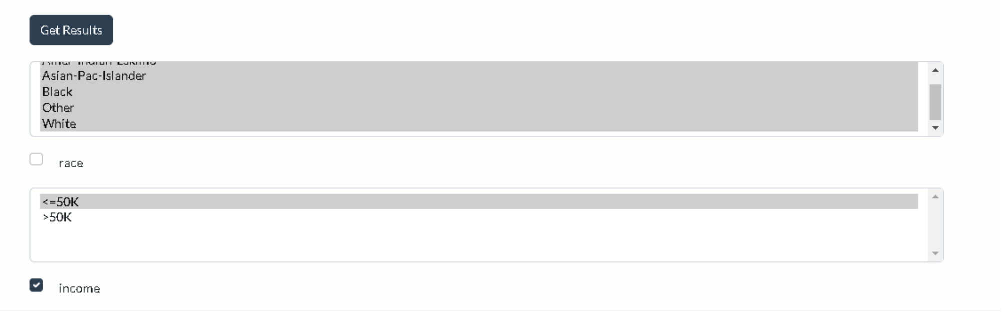
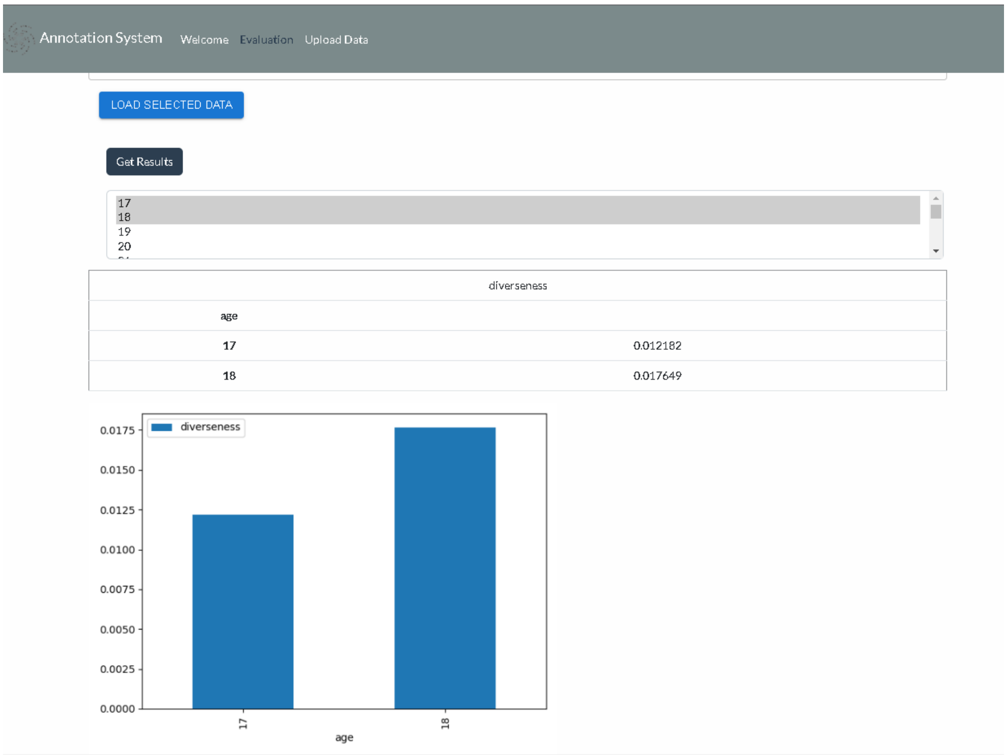


Figure 4.9: Select desirable values for calculations and calculate

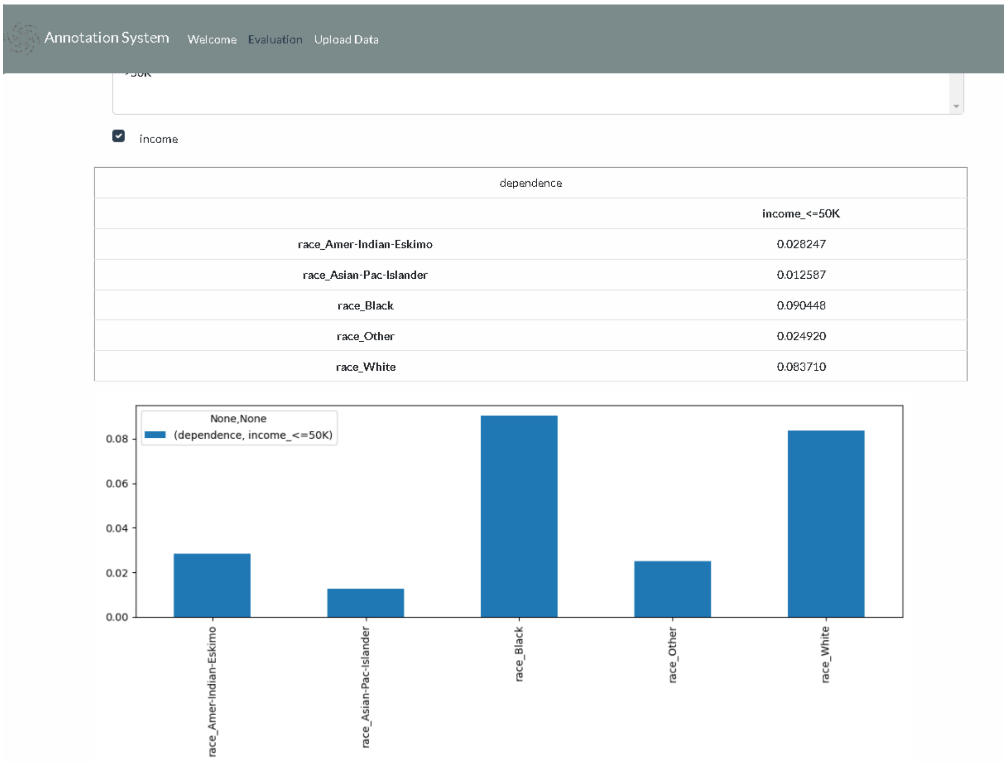
Now after selection of the values and options for input it is possible to run and get the results by results button as it is shown in the figure 4.10



diverseness	
age	
17	0.012182
18	0.017649

Figure 4.10: Running by get results button after loading data

Another type of input which is an input checkbox called income is shown in the figure 4.11 which is provided to calculate dependence of values in a dataset.



☑ income

dependence	
	income_<=50K
race_Amer-Indian-Eskimo	0.028247
race_Asian-Pac-Islander	0.012587
race_Black	0.090448
race_Other	0.024920
race_White	0.083710

Figure 4.11: Depenence sample results after running

4.2 Current Configurations

Currently there are 4 configurations for calculation of different modules. These modules are currently Dependence, Diverseness, Inclusiveness, Training Likelihood.

4.2.1 Dependence evaluation

Dependence is the relationship between two or more variables in a dataset in which the value of one variable influences or affects the value of another variable. Dependency in a dataset means that the variables are not independent and that the values of one variable can provide information about the values of another.

Consider a dataset of people's height and weight, for example. Height and weight are presumably linked in the sense that taller individuals tend to weigh more. In this instance, both the height and weight variables are dependent on one another. This relationship can be depicted using a scatterplot, where each point represents an individual and the x- and y-axes represent the height and weight, respectively.

Dependency in a dataset can alter the precision of statistical models and the interpretation of results. For instance, if a statistical model is created to predict a person's weight based on their height, the model may be biased if the variables of height and weight are highly interdependent. To address this issue, it may be required to employ more intricate models or to modify the variables in some way so that they are more independent.

There are several ways to calculate correlation in Python, and the choice of method will depend on the specific requirements of your project. One alternative to using panda get dummies() to calculate correlation is to use the corr() method provided by the pandas library. The corr() method can be applied to a DataFrame or a Series, and it calculates the Pearson correlation coefficient between the different columns or elements.

It is possible also use the scipy.stats.pearsonr() method to calculate the correlation coefficient and p-value for a set of data. This method can calculate the correlation between specific values instead of all the values in the dataframe.

Another alternative is to use the numpy.corrcoef() method, which calculates the correlation matrix of a set of variables. This method can also be used to calculate the correlation between specific values in a DataFrame or a Series.

Therefore, It is possible to use any of the above methods as per project use case and requirements. This will give you a Value Error, as the 'b' column contains string values. In order to calculate the correlation between 'a' and 'b' you should encode the column 'b' to numerical values. it is possible to use get dummies to do this in order to follow the equasions to calculate dependency:

$$\begin{gathered}
C(x_i, y_i) = f(x_i, y_i) - f'(x_i, y_i) \\
C = \sqrt{\frac{X^2}{X^2 + n}} \\
w = \sqrt{\sum_{i=1} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}} \\
w = \sqrt{\frac{C^2}{1 - C^2}} \\
x^2 = \sum_{i,j} \frac{C^2(x_i, y_i)}{n_{i,j}}
\end{gathered} \tag{4.1}$$

4.2.2 Diverseness measurement

Estimating the diversity of a dataset involves determining the number of distinct items, classes, or categories present in a dataset. This estimator aims to determine the data's richness and diversity, as well as any biases or imbalances in the distribution of items. Diverse statistical metrics, such as the frequency of occurrence, entropy, Gini index, and Simpson's index, can be used to measure diversity. The selection of a metric is dependent upon the type of data and the particular use case. Greater diversity in a dataset can result in more robust models and a more accurate depiction of the underlying population. We are able to utilize equation (4.2) to calculate the probability in order to obtain diversity measurement and python code for this can be found in the configuration file diversity.txt contained within the application in directory of formulas.

$$P = \frac{number of favorable cases}{number of possible cases} \tag{4.2}$$

4.2.3 Inclusiveness assessment

In a dataset, inclusiveness refers to the representation of a varied range of organizations and individuals, including those that are often marginalized or underrepresented. A comprehensive dataset takes into account gender, race, ethnicity, age, aptitude, socioeconomic level, and other characteristics of diversity.

For instance, a dataset of job applications may be inclusive if it includes a varied range of people from various backgrounds, such as individuals with disabilities, those from underrepresented racial and ethnic groups, and individuals from various socioeconomic backgrounds. The purpose of inclusivity in a dataset is to guarantee that the data accurately represent the experiences and viewpoints of a wide variety of people.

$$\begin{aligned} P(A = a \cap B = b) = P(A = a)P(B = b|A = a) \\ P(B = b \cap A = a) = P(B = b)P(A = a|B = b) \end{aligned} \tag{4.3}$$

4.2.4 Training Likelihood evaluation

In machine learning,to evaluate the model, the probability of the observed data given the model parameters is computed. The likelihood represents the probability of observing the data given the model's parameters.Maximum Likelihood Estimation (MLE) identifies the parameters and that maximize the likelihood function, such that the model provides the highest probability for the observed data. The training likelihood can be used to compare and pick the best model among several alternatives. A greater training likelihood value suggests that the model is a good fit for the data. It is essential to keep in mind, however, that the training likelihood is merely a measure of how well the model fits the training data and may not generalize well to unknown data. Overfitting, in which a model fits the training data too well but does not generalize well to new data, is a typical issue when the training likelihood is used as the sole criterion for evaluation. Cross-validation and other techniques that evaluate the model's performance using held-out data are employed to address this issue. Evaluation of training likelihood measures how well a probabilistic model matches the observed data. In a probabilistic model, we make assumptions about the data's underlying generative process and use a set of parameters to model it. Finding the parameters that best explain the observed data is the purpose. Using the euqation (4.4) training likelihood can be determined. regarding the computation of inclusiveness we are able to mention that from the coding standpoint, we are using the intersection value from the equations in (4.3) dividing it by probability of target value.

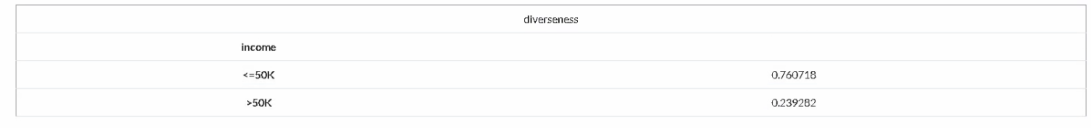
$$\begin{aligned} P(A = a|B = b) &= \frac{P(A = a)P(B = b|A = a)}{P(B = b)} \\ P(B = b \cap A = a) &= \frac{P(B = b)P(A = a|B = b)}{P(A = a)} \end{aligned} \tag{4.4}$$

4.3 Results

After executing an application over three distinct initial datasets containing specified values, the following outcomes are observed. Three datasets, namely Adults, Compass, and Drugs, are utilized to generate these results.

4.3.1 Results of adults dataset

The adult dataset [9] is an extensively used machine learning dataset that is also known as the Census Income dataset. It includes information about individuals, such as their age, education, employment status, income, etc. The dataset is used for binary classification tasks in which the objective is to predict whether an individual's annual income is greater than or less than $50,000 based on the other features. It is commonly employed for the practice and evaluation of machine learning algorithms. Here is the results of diverseness module for adults dataset as table 4.12 and chart 4.13.



	diverseness
income	
<=50K	0.760718
>50K	0.239282

Figure 4.12: Adults income diverseness

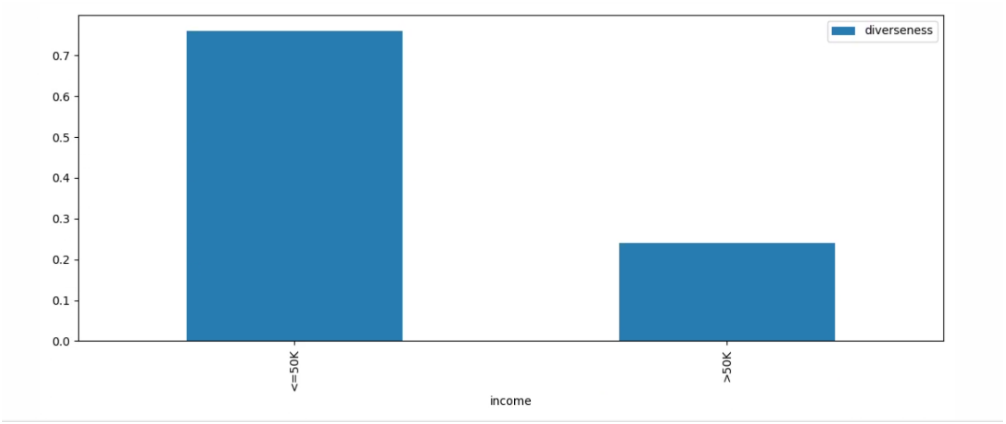
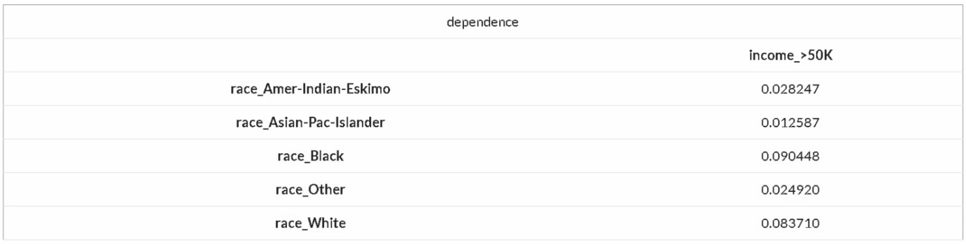


Figure 4.13: Adults income diverseness chart

Here is the results of dependence module for adults dataset as table 4.14 and chart 4.15.



dependence	
	income_>50K
race_Amer-Indian-Eskimo	0.028247
race_Asian-Pac-Islander	0.012587
race_Black	0.090448
race_Other	0.024920
race_White	0.083710

Figure 4.14: Adults race income dependence

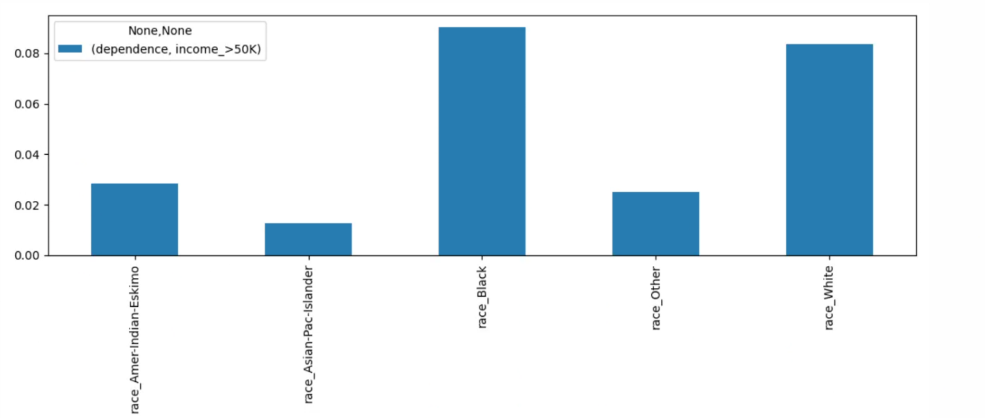
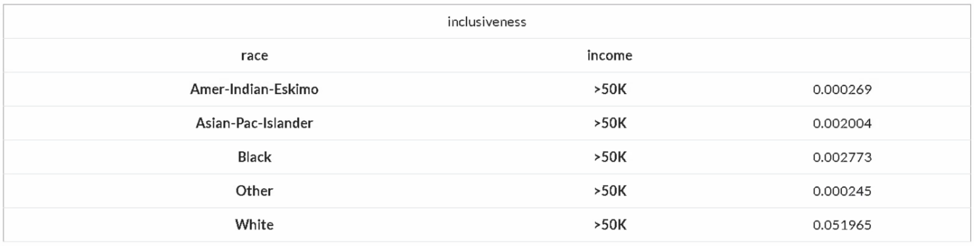


Figure 4.15: Adults race income dependence Chart

The results of inclusiveness module for adults dataset are available as table 4.16 and chart 4.17.



inclusiveness		
race	income	
Amer-Indian-Eskimo	>50K	0.000269
Asian-Pac-Islander	>50K	0.002004
Black	>50K	0.002773
Other	>50K	0.000245
White	>50K	0.051965

Figure 4.16: Adults race income inclusiveness

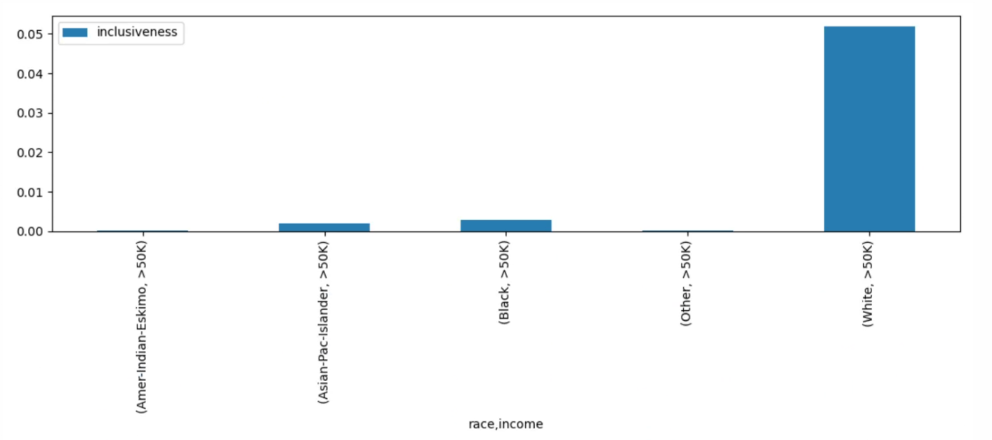
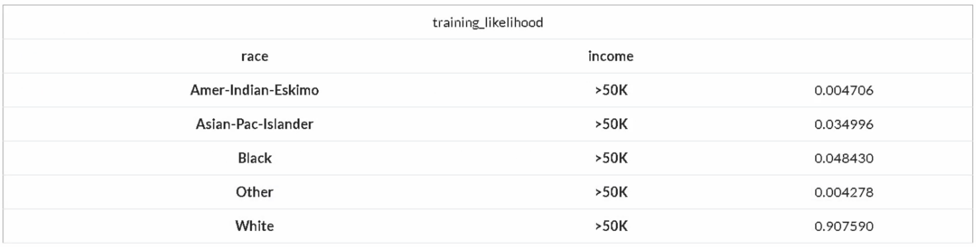


Figure 4.17: Adults race income inclusiveness chart

Here it is possible to find the results of likelihood module for adults dataset as table 4.18 and chart 4.19.



training_likelihood		
race	income	
Amer-Indian-Eskimo	>50K	0.004706
Asian-Pac-Islander	>50K	0.034996
Black	>50K	0.048430
Other	>50K	0.004278
White	>50K	0.907590

Figure 4.18: Adults race income likelihood

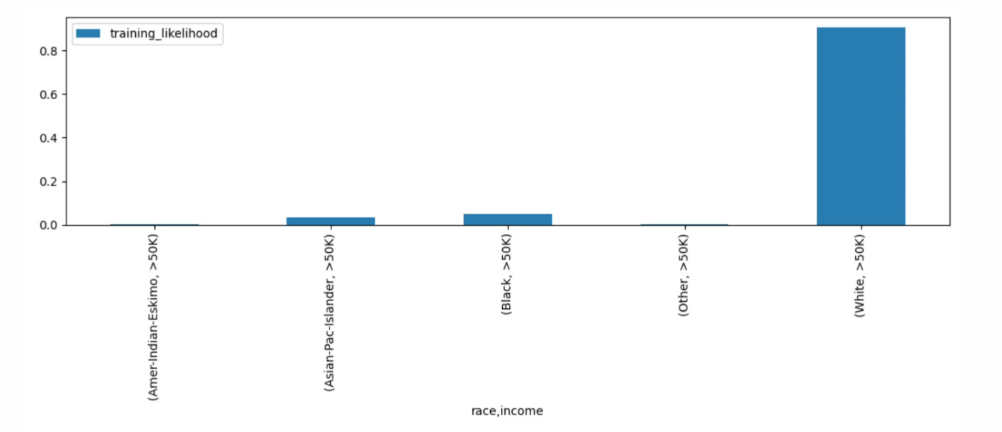
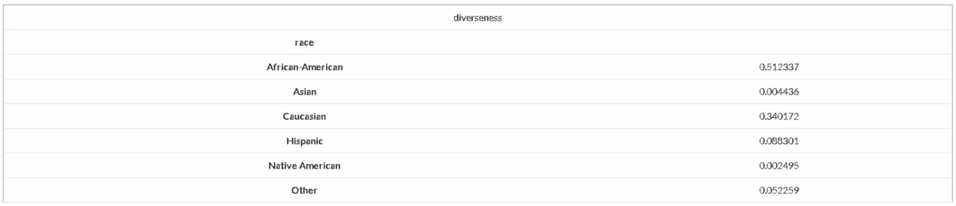


Figure 4.19: Adults race income likelihood chart

4.3.2 Results of compass dataset

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)[8] is a collection of data pertaining to persons who have been evaluated by the Northpointe COMPAS risk assessment program. In the U.S. criminal justice system, the instrument is used to predict the likelihood of recidivism among defendants. The dataset includes demographic data, criminal history, and various risk scores issued by the COMPAS system. The use of COMPAS and similar technologies in the criminal justice system has been the topic of controversy and debate, since some have raised concerns about the possibility of bias in the forecasts and the influence these projections may have on the future of an individual.

Here is a table 4.20 and chart 4.21 presenting the results of the diversity module for the COMPAS dataset.



	diverseness
race	
African-American	0.512337
Asian	0.004436
Caucasian	0.340172
Hispanic	0.088301
Native American	0.002495
Other	0.052259

Figure 4.20: COMPAS race diverseness

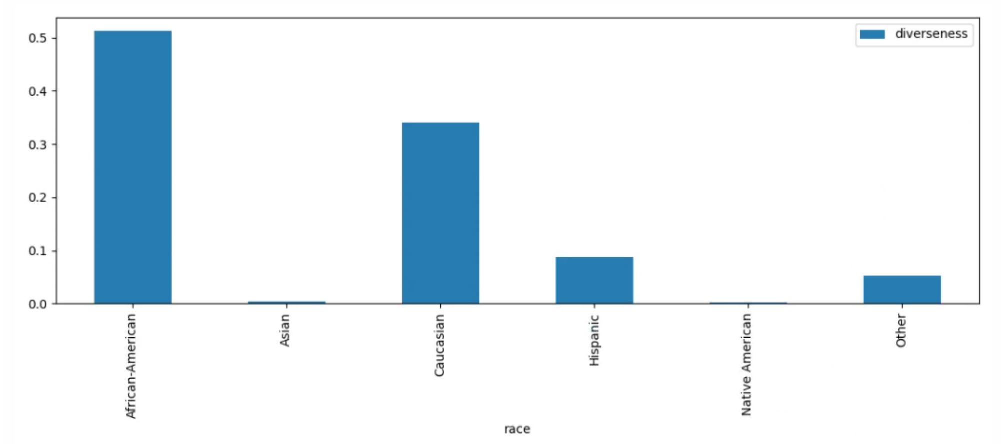
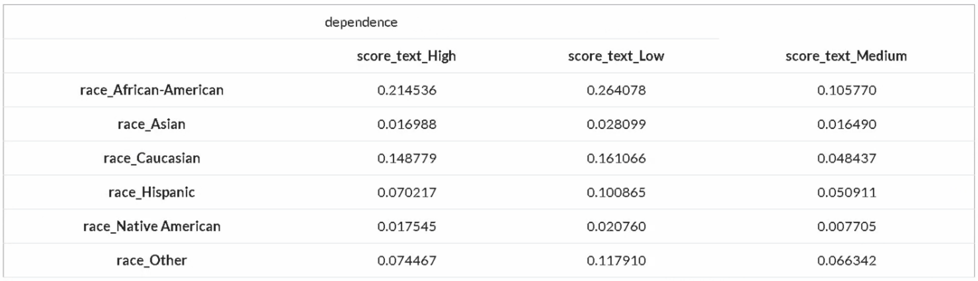


Figure 4.21: COMPAS race diverseness chart

Table 4.22 and chart 4.23 showcasing the outcome of the dependence module for the COMPAS dataset is provided here.



	dependence		
	score_text_High	score_text_Low	score_text_Medium
race_African-American	0.214536	0.264078	0.105770
race_Asian	0.016988	0.028099	0.016490
race_Caucasian	0.148779	0.161066	0.048437
race_Hispanic	0.070217	0.100865	0.050911
race_Native American	0.017545	0.020760	0.007705
race_Other	0.074467	0.117910	0.066342

Figure 4.22: COMPAS race score dependence

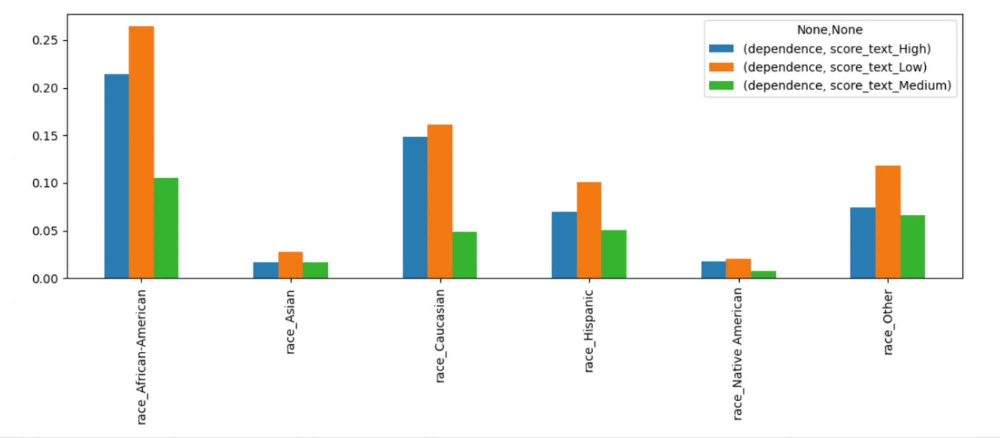
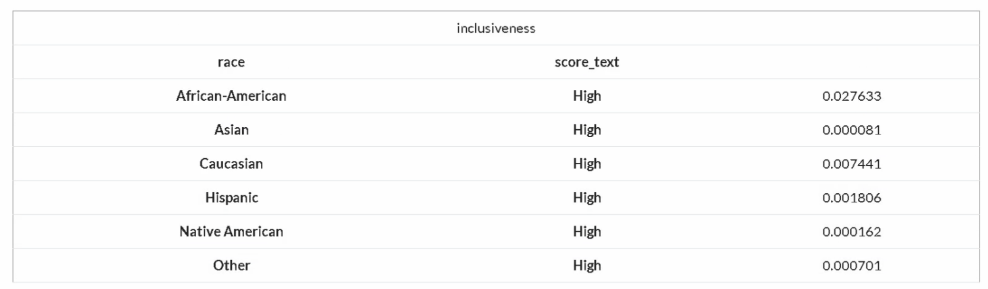


Figure 4.23: COMPAS race score dependence Chart

This is the table 4.24 and chart 4.25 presenting the results obtained from the inclusiveness module applied to the COMPAS dataset.



	inclusiveness	
race	score_text	
African-American	High	0.027633
Asian	High	0.000081
Caucasian	High	0.007441
Hispanic	High	0.001806
Native American	High	0.000162
Other	High	0.000701

Figure 4.24: COMPAS race score inclusiveness

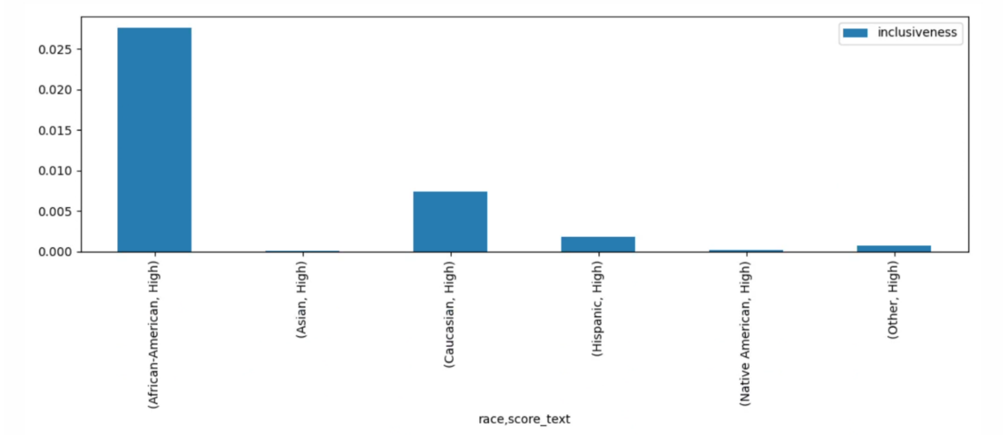
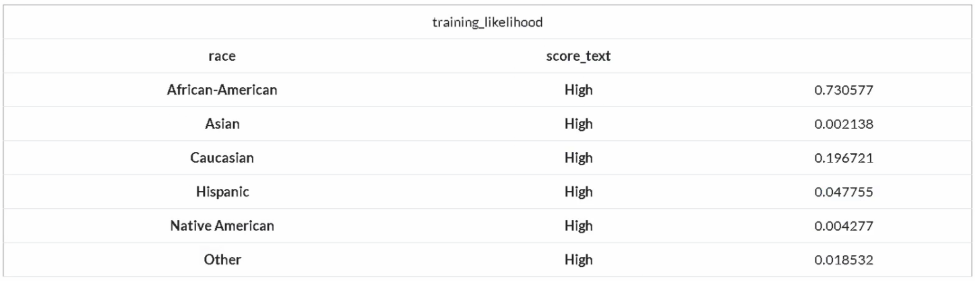


Figure 4.25: COMPAS race score inclusiveness chart

Here is the results of likelihood module for COMPAS dataset as table 4.26 and chart 4.27.



	training_likelihood	
race	score_text	
African-American	High	0.730577
Asian	High	0.002138
Caucasian	High	0.196721
Hispanic	High	0.047755
Native American	High	0.004277
Other	High	0.018532

Figure 4.26: COMPAS race score likelihood

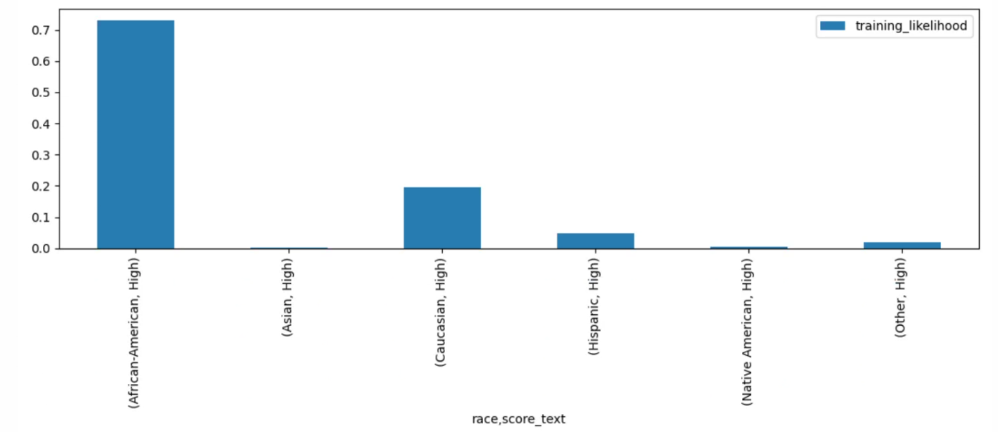
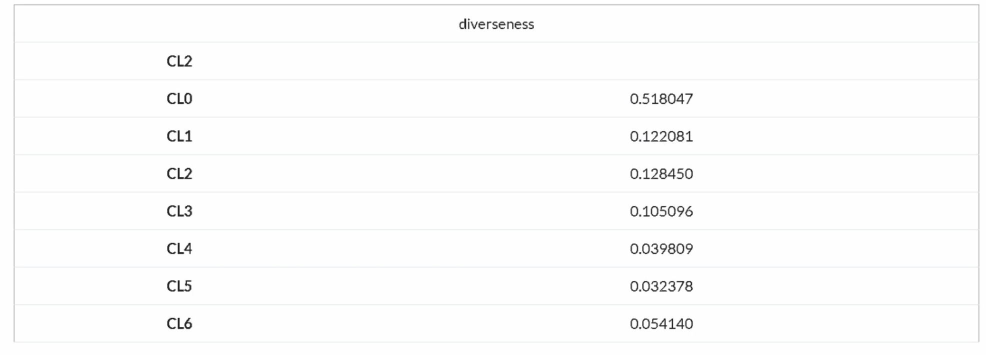


Figure 4.27: COMPAS race score likelihood Chart

4.3.3 Results of drug dataset

The "Drugs" dataset[9],[10] is widely utilized in machine learning, particularly in the fields of pharmacology and drug discovery. It often comprises information about numerous pharmaceuticals, their chemical structures, and their varied qualities, such as activity against specific diseases, adverse effects, and drug interactions. The dataset is utilized for tasks such as predicting drug-target interactions, drug efficacy, and drug repositioning. This dataset can be used to train machine learning algorithms to predict various drug properties and aid in drug discovery and development.

Here is the results of diverseness module for Drugs dataset as table 4.28 and chart 4.29.



	diverseness
CL2	
CL0	0.518047
CL1	0.122081
CL2	0.128450
CL3	0.105096
CL4	0.039809
CL5	0.032378
CL6	0.054140

Figure 4.28: Drugs CL2 Diversness

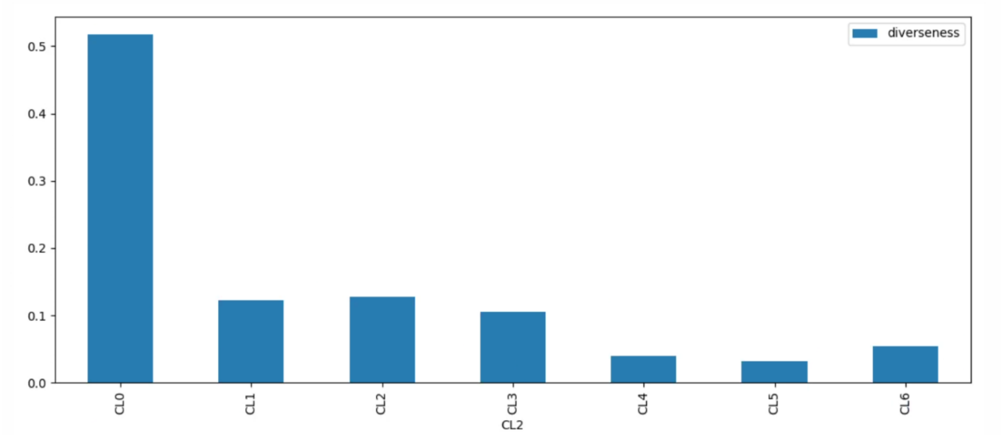
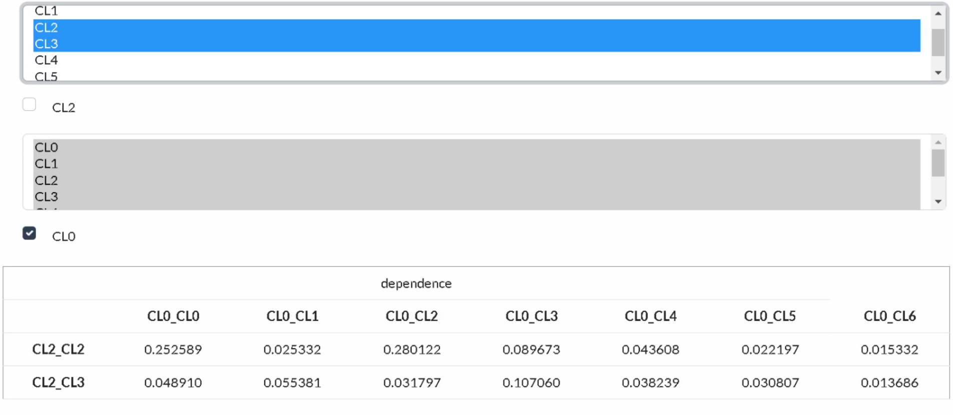


Figure 4.29: Drugs CL2 CL0 Diversness Chart

Here is the results of dependence module for Drugs dataset as table 4.30 and chart 4.31.



CL1
CL2
CL3
CL4
CL5

□ CL2

CL0
CL1
CL2
CL3

☑ CL0

dependence

	CL0_CL0	CL0_CL1	CL0_CL2	CL0_CL3	CL0_CL4	CL0_CL5	CL0_CL6
CL2_CL2	0.252589	0.025332	0.280122	0.089673	0.043608	0.022197	0.015332
CL2_CL3	0.048910	0.055381	0.031797	0.107060	0.038239	0.030807	0.013686

Figure 4.30: Drugs CL2 CL0 dependence

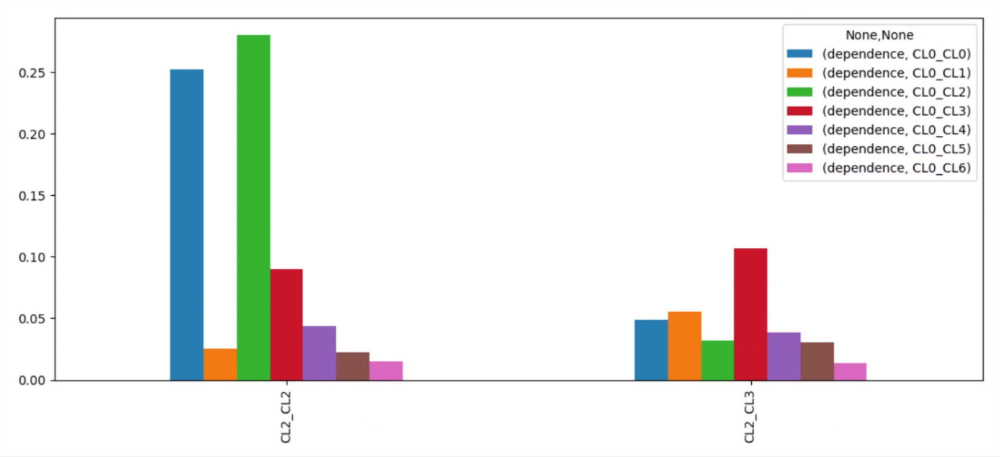
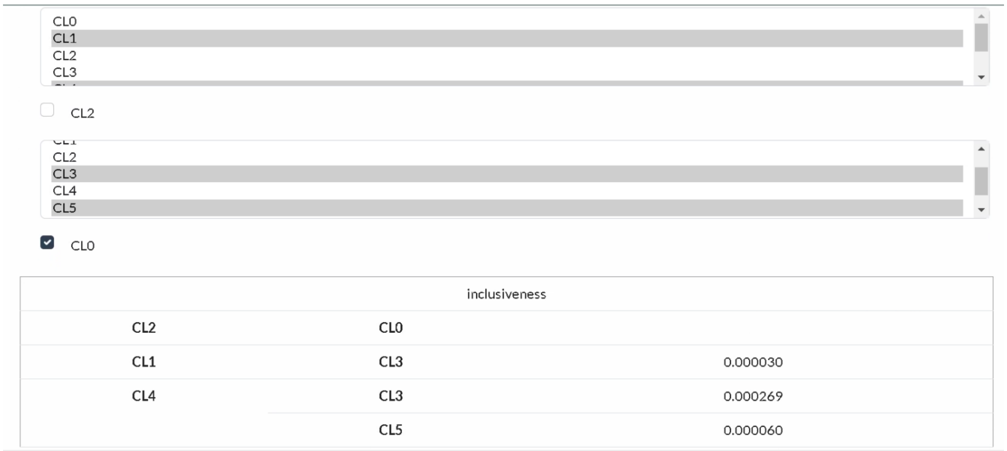


Figure 4.31: Drugs CL2 CL0 dependence chart

The output of the inclussiveness module for the Drugs dataset has been arranged and presented into the table 4.32 and chart 4.33.



□ CL2

☑ CL0

		inclusiveness
CL2	CL0	
CL1	CL3	0.000030
CL4	CL3	0.000269
	CL5	0.000060

Figure 4.32: Drugs CL2 CL0 inclusiveness

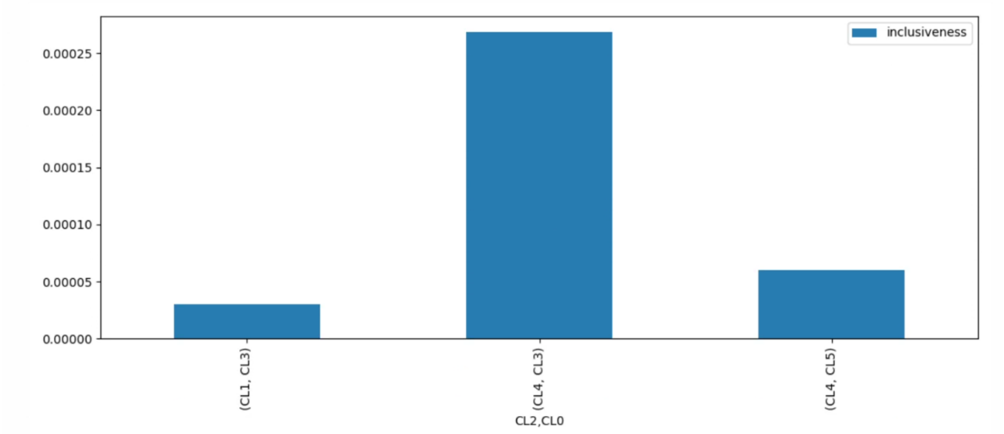
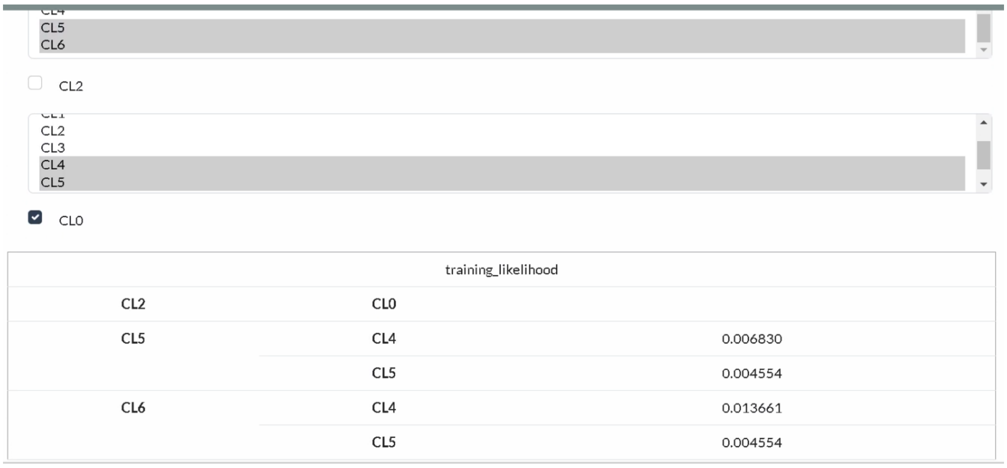


Figure 4.33: Drugs CL2 CL0 inclusiveness chart

The findings from the likelihood module for the Drugs dataset are presented in the table 4.34 and chart 4.35. format here.



training_likelihood

CL2	CL0	
CL5	CL4	0.006830
	CL5	0.004554
CL6	CL4	0.013661
	CL5	0.004554

Figure 4.34: Drugs CL2 CL0 likelihood

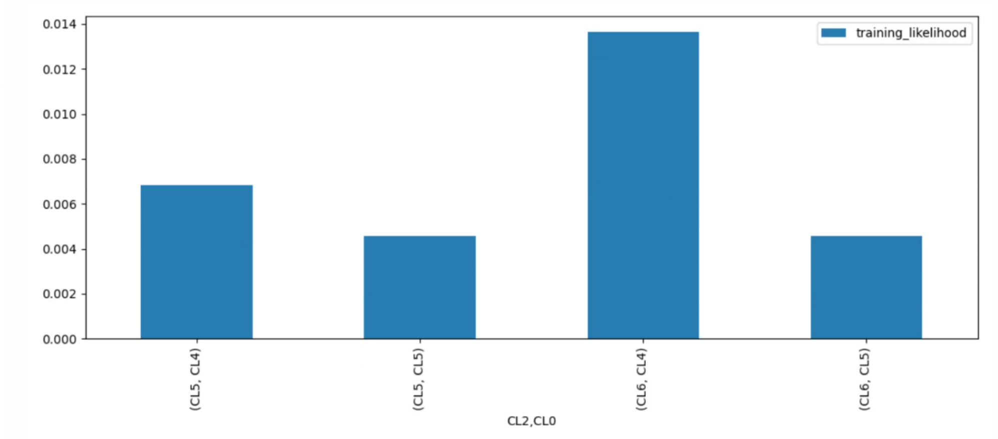


Figure 4.35: Drugs CL2 CL0 likelihood chart

Chapter 5

Conclusions and potential future development

This application currently have 4 modules with specified configurations. As the core of application serves input formulas files in order to get values and compute the results, it is easy to add more modules in the future in order to have more possibilities in calculations.For instance using Gini coeficient.Gini coefficient is a value between 0 and 1 that measures inequality in a data set.A number of 0 denotes complete equality (i.e., all values are identical), whereas a value of 1 denotes maximum inequality (i.e., one value accounts for 100% of the total). It can be used to calculate the total amount of income or wealth for the population being considered in a dataset containing certain users.

To expand our system and make it more comprehensive, you could consider incorporating additional modules that address other aspects of discriminatory risk. Here are a few suggestions:
1. Bias Detection: You could add a module that specifically identifies sources of bias in the data and the annotations. This could involve techniques such as fairness and accountability analysis.
2. Human in the Loop: You could add a human review component to your system to provide additional oversight and ensure that the decisions made by the system are ethical and fair. This could involve a combination of human annotators and machine learning algorithms working together.
3. Explanation Generation: You could add a module that generates explanations for the decisions made by the system. This could help users understand why the system is making certain predictions and could also provide valuable insights for future improvement.
4. Privacy and Confidentiality: You could add a module that ensures that sensitive information is protected and that the privacy of individuals is maintained. This

could involve techniques such as differential privacy or secure multiparty computation.
5. Adversarial Robustness: You could add a module that makes the system more robust against adversarial attacks. This could involve techniques such as adversarial training or robust optimization.
some specific measures could be considered adding to our system to detect discriminatory risk:

1. Gini Coefficient: This measure is used to quantify the inequality in a data set, and could be used to assess the diversity and inclusiveness of your data and annotations.A value of 0 expressing total equality and a value of 1 maximal inequality. It is commonly used as a measure of inequality of income or wealth distribution. The Gini coefficient is defined as the ratio of the area between the Lorenz curve (a graphical representation of the cumulative distribution of wealth or income) and the line of equality to the total area under the line of equality[30]. A higher Gini coefficient indicates a higher level of inequality in the distribution. The Gini coefficient is widely used as a summary statistic for inequality, and as a measure of the dispersion of a frequency distribution.

$$G = \frac{1}{2n} \frac{\sum |x_i - x_j|}{\sum x_i x_j} \tag{5.1}$$

   where n is the number of values in the distribution, x_i and x_j are two values in the distribution, and the sum is taken over all possible pairs of values $(i \neq j)$. This formula can be used to calculate the Gini coefficient for a given set of values representing a distribution.

2. Entropy Measures: Entropy measures can be used to quantify the randomness or unpredictability of a data set like Shannon entropy [31].

$$H(S) = -\sum (p(i) \times \log_2 p(i)) \tag{5.2}$$

   This can be useful in detecting discriminatory risk by identifying patterns or correlations in the data that could result in discriminatory outcomes.

3. Precision, Recall, and F1 Score: These measures can be used to evaluate the accuracy of your system's predictions and can help you identify areas for improvement. You could use these measures to ensure that your system is making predictions that are both accurate and inclusive.

Precision, recall, and F1 score are evaluation metrics used to measure the performance of a binary classifier.

Precision measures the proportion of positive predictions that are actually positive. Precision is defined as:

$$Precision = \frac{TruePositives}{(TruePositives + FalsePositives)} \tag{5.3}$$

where True Positives (TP) are the number of samples that are correctly classified as positive, and False Positives (FP) are the number of samples that are incorrectly classified as positive.

Recall, also known as sensitivity or the true positive rate, measures the proportion of positive samples that are correctly classified as positive. Recall is defined as:

$$Recall = \frac{TruePositives}{(TruePositives + FalseNegatives)} \tag{5.4}$$

where False Negatives (FN) are the number of samples that are incorrectly classified as negative.

The F1 score is the harmonic mean of precision and recall and is defined as:

$$F1Score = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \tag{5.5}$$

The F1 score provides a single number that balances precision and recall. A classifier with high precision and low recall will have a lower F1 score than a classifier with low precision and high recall, making the F1 score a useful metric for comparing classifiers with imbalanced precision and recall.

4. Demographic Parity: This measure assesses the equality of treatment across different groups of individuals. You could use this measure to ensure that your system is treating all individuals fairly, regardless of their demographic characteristics.

   Demographic parity is a metric used to measure the fairness of a binary classification system with respect to a protected attribute, such as race or gender. Demographic parity requires that the positive prediction rate (i.e., the proportion of samples predicted as positive) be equal across different groups defined by the protected attribute.

To calculate demographic parity in a dataset, you first need to determine the positive prediction rate for each group defined by the protected attribute. Given a binary classification system with protected attribute A and classes C1 (positive) and C2 (negative), the positive prediction rate for group g defined by attribute A can be calculated as[32]:

$$PPV(g) = \frac{\text{number of samples in group g predicted as C1}}{\text{total number of samples in group g}} \tag{5.6}$$

Once you have calculated the positive prediction rate for each group, you can calculate demographic parity as the absolute difference between the highest and lowest positive prediction rates:

$$DP = |PPV(g_1) - PPV(g_2)| \tag{5.7}$$

where g1 and g2 are two groups defined by the protected attribute A. A value of 0 for DP indicates that demographic parity has been achieved, meaning that the positive prediction rate is equal across all groups defined by the protected attribute. A non-zero value indicates that demographic parity has not been achieved, and the magnitude of the value indicates the extent to which demographic parity has been violated.

5. Equal Opportunity: This measure assesses the fairness of positive outcomes, such as approval or acceptance, across different groups of individuals. You could use this measure to ensure that your system is providing equal opportunities for positive outcomes to all individuals. The metric is defined as[33]:

$$\begin{aligned}\text{Equal Opportunity} = |&\text{True Positive Rate for Group A}- \\ &\text{True Positive Rate for Group B}|\end{aligned} \tag{5.8}$$

where Group A and Group B are two different groups based on the protected attribute, and the True Positive Rate (TPR) for each group is calculated using Recall(5.4) equation in previous measure.

For future improvements non-functional requirement NFR1, which was discussed in the previous chapter in Table.3.2, can be utilized to enhance the system's functionalities.

Bibliography

[1] Amir E. Khandani, Adlar J. Kim, and Andrew Lo. «Consumer credit-risk models via machine-learning algorithms». In: *Journal of Banking Finance* 34.11 (2010), pp. 2767–2787 (cit. on p. 1).

[2] Amir E. Khandani, Adlar J. Kim, and Andrew Lo. «Predicting crime using Twitter and kernel density estimation». In: *Decision Support Systems* 61.11 (2014), pp. 115–125. ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2014.02.003. URL: https://www.sciencedirect.com/science/article/pii/S0167923614000268 (cit. on p. 1).

[3] Abdul Hamid M. Ragab, Amin Y. Noaman, Abdullah S. Al-Ghamdi, and Ayman I. Madbouly. «A Comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data Mining». In: 2014, pp. 106–113. DOI: 10.1145/26 (cit. on p. 1).

[4] *Visual analysis of discrimination in machine learning*. https://deepai.org/publication/visual-analysis-of-discrimination-in-machine-learning. Accessed: 2023-02-01 (cit. on p. 1).

[5] *Data analyst interview questions and answers most frequently asked*. https://onlineyukti.com/data-analyst-interview-questions-and-answers-most-frequently-asked/. Accessed: 2023-02-01 (cit. on p. 1).

[6] *Gender equality in recruitment and career progression*. https://eige.europa.eu/gender-mainstreaming/toolkits/gear/gender-equality-recruitment-and-career-progression. Accessed: 2023-02-13 (cit. on p. 2).

[7] *bayesian-inference*. https://medium.com/@skilltohire/bayesian-inference-in-ad2cf534c8f1. Accessed: 2023-02-01 (cit. on p. 2).

[8] *Jeff Harry Thornburg Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. ProPublica. Retrieved September 2, 2020.* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed: 2023-02-01 (cit. on pp. 3, 14, 34).

[9] *Elaine Fehrman, Vincent Egan, and Evgeny M. Mirkes. 2015. UCI Machine Learning Repository.* http://archive.ics.uci.edu/ml. Accessed: 2023-02-01 (cit. on pp. 3, 14, 31, 38).

[10] *Elaine Fehrman, Awaz K. Muhammad, Evgeny M. Mirkes, Vincent Egan, and Alexander N. Gorban. 2017. The Five Factor Model of Personality and Evaluation of Drug Consumption Risk. , 231–242 pages.* https://doi.org/10.1007/978-3-319-55723-6_18. Accessed: 2023-02-01 (cit. on pp. 3, 14, 38).

[11] *COMPAS Recidivism Racial Bias.* https://www.kaggle.com/datasets/danofer/compass. Accessed: 2023-02-13 (cit. on p. 3).

[12] Kyra Yee, Uthaipon Tantipongpipat, and Shubhanshu Mishra. «Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency». In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (Oct. 2021), pp. 1–24. DOI: 10.1145/3479594. URL: https://doi.org/10.1145%2F3479594 (cit. on p. 6).

[13] Masatomo Yoshida and Masahiro Okuda. «Adversarial Examples for Image Cropping in Social Media». In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 2022, pp. 4898–4902. DOI: 10.1109/ICASSP43922.2022.9746949 (cit. on p. 6).

[14] *Solving the problem of racially discriminatory advertising on Facebook.* https://www.brookings.edu/research/solving-the-problem-of-racially-discriminatory-advertising-on-facebook/. Accessed: 2023-02-13 (cit. on p. 6).

[15] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. «Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes». In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: 10.1145/3359301. URL: https://doi.org/10.1145/3359301 (cit. on p. 7).

[16] David Maloney, Sung-Chul Hong, and Barin N. Nag. «Two class Bayes point machines in repayment prediction of low credit borrowers». In: *Heliyon* 8.11 (2022), e11479. ISSN: 2405-8440. DOI: https://doi.org/10.1016/j.heliyon.2022.e11479. URL: https://www.sciencedirect.com/science/article/pii/S2405844022027670 (cit. on p. 7).

[17] Fan Liu, Zhongsheng Hua, and Andrew Lim. «Identifying future defaulters: A hierarchical Bayesian method». In: *European Journal of Operational Research* 241.1 (2015), pp. 202–211. ISSN: 0377-2217. DOI: https://doi.org/10.1016/j.ejor.2014.08.008. URL: https://www.sciencedirect.com/science/article/pii/S0377221714006432 (cit. on p. 7).

[18] Elena Beretta, Antonio Vetro, Bruno Lepri, and Juan Carlos De Martin. «Detecting discriminatory risk through data annotation based on Bayesian inferences». In: Mar. 2021, pp. 794–804. DOI: 10.1145/3442188.3445940 (cit. on pp. 7, 10, 14).

[19] Marco Torchiano Mariachiara Mecati Antonio Vetrò. «Detecting Discrimination Risk in Automated Decision-Making Systems with Balance Measures on Input Data». In: Jan. 2022, pp. 4287–4296. DOI: 10.1109/BigData52589.2021.9671443 (cit. on p. 7).

[20] Raffaele Giancotti, Patrizia Vizza, Giuseppe Tradigo, and Pierangelo Veltri. «A framework for clinical data integration and annotation for decision support». In: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2021, pp. 3018–3020. DOI: 10.1109/BIBM52615.2021.9669611 (cit. on p. 7).

[21] Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. «Comparing Bayesian Models of Annotation». In: *Transactions of the Association for Computational Linguistics* 6 (Dec. 2018), pp. 571–585. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00040. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00040/1567662/tacl_a_00040.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00040 (cit. on p. 7).

[22] *What Is ISO 25010?* https://www.perforce.com/blog/qac/what-is-iso-25010. Accessed: 2023-02-01 (cit. on pp. 10, 13).

[23] *Imbalanced data as risk factor of discriminating automated decisions: a measurement-based approach.* https://www.jipitec.eu/issues/jipitec-12-4-2021/5452. Accessed: 2023-02-01 (cit. on pp. 11, 12, 14).

[24] *R Vs Python: What's the Difference?* https://www.guru99.com/r-vs-python.html. Accessed: 2023-02-01 (cit. on p. 11).

[25] *R-vs-Python.* https://www.simplilearn.com/r-vs-python-battle-of-programming-languages-article. Accessed: 2023-02-01 (cit. on pp. 11, 12).

[26] *Is Python faster than R.* https://towardsdatascience.com/is-python-faster-than-r-db06c5be5ce8. Accessed: 2023-02-01 (cit. on p. 11).

[27] *Django and matplotlib integration | How to use matplotlib with Django.* https://www.youtube.com/watch?v=jrT6NiM46jk. Accessed: 2023-02-01 (cit. on p. 12).

[28] *The War on Data Science: Python versus R.* https://hub.packtpub.com/war-data-science-python-versus-r/. Accessed: 2023-02-01 (cit. on p. 12).

[29] *Visualizzazione dell'Informazione Quantitativa.* https://softeng.polito.it/courses/VIQ/. Accessed: 2023-02-01 (cit. on p. 14).

[30] Anthony B Atkinson and Francois Bourguignon, eds. *Handbook of Income Distribution.* Elsevier Science, 2000 (cit. on p. 44).

[31] Matthew A. Smith Fernando Montani Adam Kohn and Simon R. Schultz. «The Role of Correlations in Direction and Contrast Coding in the Primary Visual Cortex (SupplementalMaterial)». In: *Journal of Neuroscience* 27.9 (2007), pp. 2338–2348. DOI: `https://doi.org/10.1523/JNEUROSCI.3417-06.2007` (cit. on p. 44).

[32] Michael Feldman, Adam Kapelner, Gavin Saxton, and John Gilbert. «Certifying and removing disparate impact». In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* The concept of demographic parity is central to the discussion of removing disparate. ACM. 2015, pp. 259–268 (cit. on p. 46).

[33] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Max Welling. «Fairness Constraints: Mechanisms for Fair Classification». In: *IEEE Transactions on Knowledge and Data Engineering* 29.8 (2017), pp. 1875–1888 (cit. on p. 46).