Politecnico di Torino

Master of Science in Biomedical Engineering

CONSTRUCTION OF AN AI-BASED SYSTEM FOR THE DETECTION OF PROSTATE CANCER

Author: Laura Lopera Tobón

Supervisors:

Prof. Gabriella Balestra

Prof. Samanta Rosati

Academic year: 2022/2023

Acknowledgements

First of all, I would like to thank my thesis supervisors for having guided me in this final phase of my journey at the Politecnico di Torino. I would also like to thank God for the opportunity, and all the people who contributed, even if it was with just a small gesture, to make this dream come true.

Mainly I want to thank my parents and family, who helped me to become the woman I am today, and who reminded me on many occasions how important it was to fight for my ideals and not to let situations defeat me.

On the other hand, I want to express my gratitude to all of my friends, from those who believed in me in Colombia and encouraged me to come to Italy, to those I met in Turin. I feel really fortunate to have such amazing people in my life.

Last but not least, I want to thank my boyfriend, who has always stood by my side, supported me no matter what, and has been my number-one fan.

After 8 years of studying I can proudly say that I am becoming an engineer, and I can affirm that no matter how small or big the contribution of each of the people I have mentioned has been, I could not have done it without them. I feel and will always feel immensely grateful.

Contents

List of Tables vii

List of Figures ix

Abstract xi

1 Introduction 1
- 1.1 Prostate cancer and the aim of the project . . . 1
- 1.2 State of the art . . . 4
  - 1.2.1 Gravina et al. (2022) . . . 5
  - 1.2.2 Chen et al. (2022) . . . 5
  - 1.2.3 Wang et al. (2022) . . . 6
  - 1.2.4 Checcucci et al. (2021) . . . 7
  - 1.2.5 Yu et al. (2021) . . . 7
  - 1.2.6 Takeuchi et al. (2019) . . . 8

2 The dataset 9
- 2.1 Analysis of the dataset . . . 10
  - 2.1.1 Computation of the median . . . 11
  - 2.1.2 Computation of the mean . . . 11
  - 2.1.3 Computation of the frequency tables . . . 12
- 2.2 Verification and modification of the dataset . . . 15

3 Imputation of the variables 19
- 3.1 Boxplot . . . 21
- 3.2 Exclusion of variables and creation of new datasets . . . 21

4 Classification 23
- 4.1 kNN description . . . 23
  - 4.1.1 Gower distance . . . 24
  - 4.1.2 Hamming distance . . . 25

4.2 kNN implementation 25

5 Results and analysis **29**

6 Conclusions **37**

Bibliography **39**

Appendix **41**

List of Tables

2.1 Statistical description for continuous integer and ordinal input variables . 11
2.2 Statistical description for continuous real input variables 12
2.4 Frequency table for the variable "DRE" 13
2.5 Frequency table for the variable "DRE abnormality lobe" 13
2.3 Frequency table for the variable "TURP" 13
2.6 Frequency table for the variable "result of the I biopsy" 13
2.7 Frequency table for the variable "result of the II biopsy" 14
2.8 Frequency table for the variable "result of the III biopsy" 14
2.9 Frequency table for the variable "lesion location" 14
2.10 Frequency table for the variable "lesion side" 15
2.11 Number of missing values for each variable on the original dataset 16
2.12 Number of missing values for each variable on the modified dataset 17

3.1 Description of the different models used for the imputation of variables . 20
3.2 Variables taken into account for the boxplot 21

5.1 Performance of the global best classifier between Datset 1 and Dataset 2: the one made with Dataset 2 normalized and Hamming distance. Values are reported in percentage (%). 34
5.2 Performance of the classifiers implemented with Dataset 3. Values are reported in percentage (%). 36

List of Figures

1.1 Representation of the prostate anatomy [2]. 1
1.2 Detailed illustration of prostate areas [4]. 2
1.3 Estimated crude incidence rates in 2020 of prostate cancer, all ages [9]. . 4

5.1 Specificity values calculated for each of the classifiers. 30
5.2 Sensitivity values calculated for each of the classifiers. 31
5.3 NPV values calculated for each of the classifiers. 31
5.4 PPV values calculated for each of the classifiers 32
5.5 Balanced accuracy values calculated for each of the classifiers 33
5.6 Confusion matrixes of the classifiers done with Gower distance, original and normalized, respectively. 35
5.7 Confusion matrixes of the classifiers done with Hamming distance, original and normalized, respectively. 35

ABSTRACT

Prostate cancer is the second cancer with the highest incidence worldwide in men and is characterized by the excessive proliferation of cells in the prostate gland. Due to this incidence, this pathology's correct and timely identification has become important. Currently, the accurate diagnosis of prostate cancer is made through biopsy, however, performing a biopsy on each suspected person entails discomfort and can even trigger health problems in patients. For this reason, other methods that are based on clinical variables are seeking to be implemented, which could avoid unnecessary biopsies. In this thesis work, the main objective was to develop an artificial intelligence-based algorithm that allows the correct prediction of prostatic cancer using MATLAB. In consequence, a dataset of 1621 patients with different types of variables was used: categorical, numerical, and binary. Initially, modifications were made to these data (e.g., completeness and correctness verification, merging of variables), and after carrying out an analysis of missing values, it was decided to implement imputation of variables using k-Nearest Neighbors (kNN) and see what the influence of the classifier would be depending on its performance. The classifier chosen for cancer prediction was kNN. In addition, the influence of two more important parameters in the classifier was analyzed: the distance and the number of k. The distances evaluated were the Hamming distance, the default of MATLAB when the input is a table and which compares string sequences, and the Gower distance, which calculates different distances depending on the type of variable and for which a respective function was designed. For the number of k, the square root of the number of subjects was originally chosen, and subsequently 75%, 50%, and 25% of its initial value. The performance of the classifiers was evaluated based on some descriptive parameters which were calculated after training the predictors with a training set and validating them through k-fold cross-validation (where 10 different groups were generated). In this study, it was found that the imputation of variables did not really have a great influence on the results of the dataset, which can be explained by the fact that almost all the fields were imputed with the same value, except for a small percentage that corresponded to less than 3% of total patients. Additionally, appreciable differences were found when changing the distance: the one corresponding to Hamming managed

to have more generalized results and with less classification error (balanced accuracy of closely 71%). The combination of the parameters that gave the best results were: use of the dataset only with the patients that had not missing values (dataset 3), normalizing, using Hamming distance and chosing k as the root square of the number of patients. Finally, it could be seen that the results obtained were partially encouraging: although the classifiers did not have optimal performance, this research would help in the future for the development of better tools in the field of prostate cancer, especially when there is a dataset as heterogeneous as the one used in this project.

1 Introduction

1.1 Prostate cancer and the aim of the project

The prostate (also known as the prostate gland) is a male reproductive system organ. It is situated in front of the rectum, inside the pelvic body cavity, beneath the urine bladder. Its size is comparable to the dimensions of a walnut, but as time passes, it tends to get larger. Seminal vesicles, or glands that create a substantial portion of the fluid that constitutes semen, are located posterior to the prostate, and the urethra passes throughout the prostate, as represented in Figure 1.1 [1].

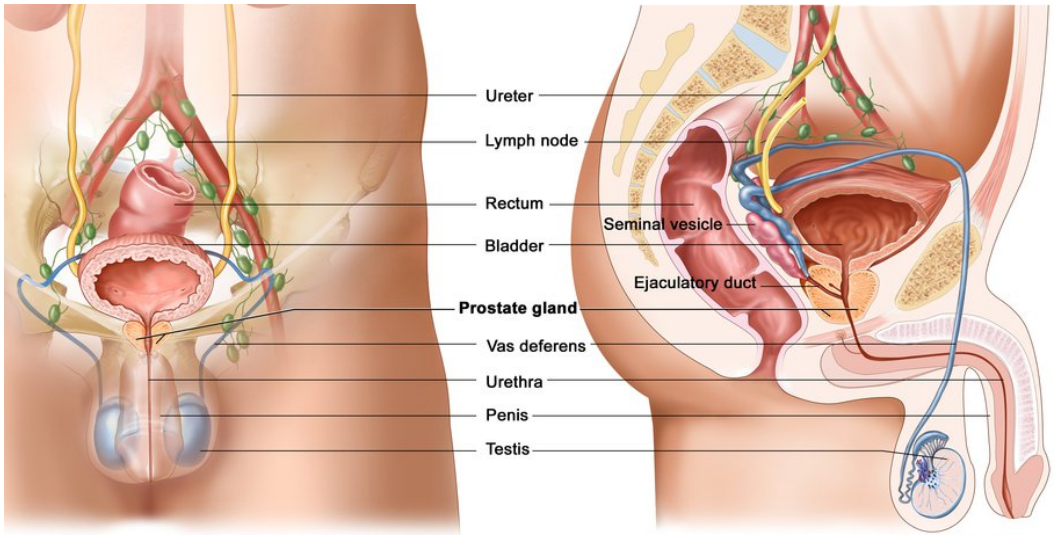


Figure 1.1. Representation of the prostate anatomy [2].

Several zones can be distinguished in the prostate, as defined by McNeal [3], which are of high clinical importance:

- **Pheripheral zone:** this zone makes up more than 70% of the glandular prostate and is where cancer most frequently spreads. Between the verumontanum, or the area where the ejaculatory ducts enter the urethra, and the urinary bladder, this area houses the proximal urethral segment of the prostate.

- **Central zone:** comprises approximately 25 % of the prostate gland. It's described as a vertical wedge of glandular tissue lateral to each ejaculatory duct with its base cephalad at the gland capsule, and it differs from the peripheral zone in terms of stroma and the glandular architecture. This zone is less predisposed to anormal growth of tissue.
- **Transitional zone:** this zone just takes up a 5% of the prostate's overall volume, it is situated in the preprostate area and lies in the convexity of the peripheral zone.

A schematic representation of this zones is shown on Figure 1.2.

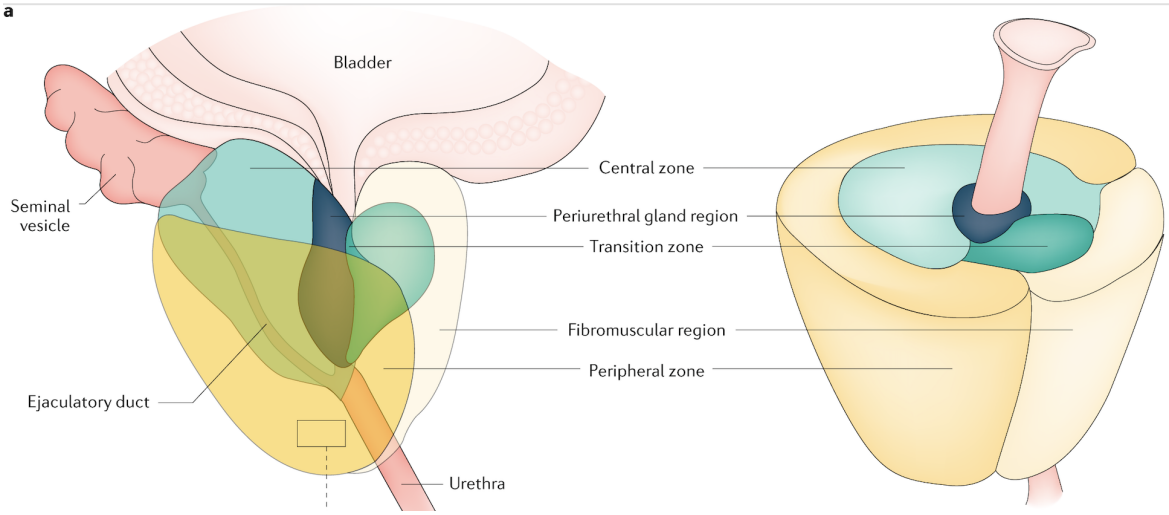


Figure 1.2. Detailed illustration of prostate areas [4].

The prostate may be exposed to a variety of benign pathologies that lead to a change in its size, including prostatitis and benign prostatic hyperplasia (BPH). Young men are more likely to get prostatitis than older men, which can be caused by bacterial or non-bacterial infections. As men age, BPH becomes increasingly prevalent in them, and even if it is not life-threatening, it can have a substantial impact on quality of life. This enlargement causes the urethra to narrow and places pressure on the base of the bladder, which produces obstruction or blockage in the flow of urine [5].

Prostate cancer, on the other hand, is a malignant condition characterized by an uncontrolled proliferation of the prostate gland's cells that results in an increase in prostatic volume. The main risk factor for this disease is age, which is typically directly correlated with cancer, i.e. the higher the age, the higher the risk. Other risk factors include familiarity, the presence of gene mutations in particular genes, and the components associated with an unhealthy lifestyle [6].

Given the possibility of multiple pathologies causing prostate gland enlargement, it is crucial to correctly diagnose the illness by considering different variables. For this

diagnosis, a rectal examination is typically carried out, which sometimes enables the presence of any nodules at the prostate level to be identified by touch, and the PSA (Prostate Specific Antigen), a protein synthesized by the cells of the prostate gland, is checked by drawing blood. The serum concentration of this protein, which is typically between 2.5 and 6.5 ng/ml, is a specific marker of the organ but not of the pathologies affecting it. To unequivocally identify the disease, it is necessary to perform a prostate biopsy, which consists of taking a piece of suspicious tissue and carrying out the relevant analyses. Before performing this biopsy, it is essential to use Multiparametric Magnetic Resonance Imaging (mpMRI) in which the lesion can be better appreciated graphically and from which various data can be taken to help make this decision [7].

When a tumor is found, it is classified via two categories: grade, which describes how aggressive the disease is, and stage, which describes how far it spreads. The so-called Gleason Score (GS), a number between 1 and 5, is given for the tumor by the pathologist who examines the tissue obtained by biopsy. It describes how similar or dissimilar the appearance of the tumor glands is from that of normal glands; the more similar they are, the lower the Gleason Score will be. The first and second most frequent ratings assigned to biopsy specimens are put together to form the Gleason Score. Low-grade tumors are those with a Gleason Score of six or less, intermediate-grade tumors are those with a score of seven, and high-grade tumors are those with a score between eight and ten. Typically, "clinically significant" cancer (csPCa) is defined as having a Gleason Score of 7 or higher. These situations raise the likelihood that the cancer will advance and spread to other organs. A new grading system has recently been created, based on the ISUP (International Society of Urological Pathology), which stratifies prostate tumors into five grades based on malignant potential and aggressiveness [8].

According to GLOBOCAN [9], prostate cancer is estimated to account for 1,414,259 new cases and 375,304 deaths (9.5% of all deaths by cancer in males) in 2020, making it the second most common malignancy in men worldwide (after lung cancer). Although only 1 in 350 males under the age of 50 will receive a prostate cancer diagnosis, the incidence rate rises to 1 in 52 men between the ages of 50 and 59. Men over the age of 65 have an incidence rate of around 60% [10]. Figure 1.3 provides an in-depth diagram of the prevalence of this disease.

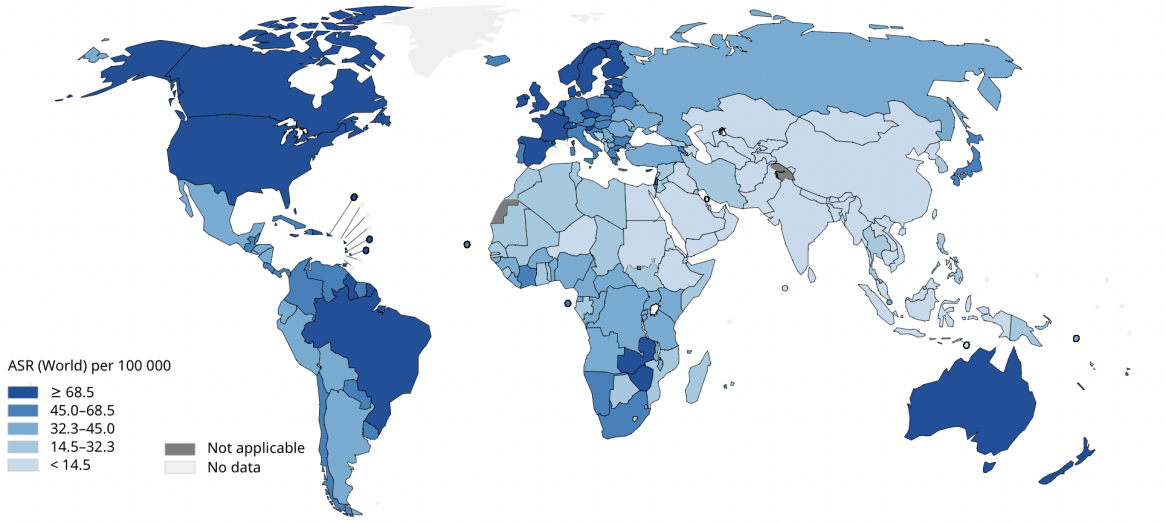


Figure 1.3. Estimated crude incidence rates in 2020 of prostate cancer, all ages [9].

The incidence has become the reason why a correct and timely diagnosis, not only of the carcinoma but also of the degree of progression of the pathology, has become indispensable. In fact, to meet this objective, over the years, a considerable number of unnecessary biopsies have been performed, which are uncomfortable for the patient and can also have complications such as infections or bleeding, especially in elderly people. To maximize the effectiveness of early detection and the clinical benefit derived, unnecessary biopsies can be avoided by constantly focusing on the development and validation of refined strategies that can establish prostate cancer risk based on noninvasively available clinical information [11]. Traditional statistical models, however, cannot easily handle the complexity of real-world problems; instead, machine learning techniques should be used because they can process the data provided to them, learn how to handle it, and automatically improve their performance.

In this thesis project, a dataset provided by the Hospital San Luigi of Orbassano which contains quite heterogeneous variables with different typologies and ranges is taken, a series of changes which include various imputations of the missing values are applied, and the parameters present in the dataset are given as input to a k-Nearest Neighbor (kNN) classifier to which some parameters are modified, to finally evaluate the performance of the different classifiers and compare the results.

1.2 State of the art

Before starting to work with the dataset, time was spent on literature research and subsequent analysis of several scientific articles related to data mining and big data. The articles were found using some free search engines for biomedical scientific literature

such as PubMed, Scopus, and Web of Science by selecting the command urology AND ('data mining' OR 'big data') as the search parameter. In this first part, 89 articles were found, and afterward, depending on the author keywords of each and the content of the abstract, 24 were chosen.

Finally, after reading the remaining articles, some were selected as important to the definition of the significative variables, others as important to the context of decision support systems, and the remaining articles were removed after realizing that there were no significant or pertinent.

1.2.1 Gravina et al. (2022)

This article by Gravina, Spirito, Celentano, *et al.* [12] focuses on the analysis of patients with a PIRADS classification of 3 because this represents a state of doubt regarding the tumor and the need for a biopsy. More specifically, different machine-learning approaches are evaluated using clinical and radiological data.

The dataset, available at the Urologic Unit of AOU Federico II in Naples, consisted of 109 patients who underwent trans-rectal prostate biopsy from January to March 2022 and included the following variables: patient weight and height, body mass index (BMI), suspect area, prostate volume, prostate-specific antigen (PSA), Psa density, free PSA, ratio, blood glucose, cholesterol, high-density lipoprotein (HDL), low-density lipoprotein (LDL), triglycerides, creatinine and variables derived from the multiparametric magnetic resonance and histological examinations.

The performance of four machine learning models was compared: classification tree (Ctree), random forest (RF), support vector machines (SVM), and feedforward neural network (NN). Random forest showed the best performance with a sensitivity that reached 81.69% and a specificity of 71.05%, resulting in a good ability to recognize the malignant class. Although the SVM and the Ctree models showed the highest sensitivity (73.68%), we chose RF as the best model as it outperformed the other models in all other metrics, whilst maintaining a good value for specificity (2.63%). The main limitation specified by the authors was the amount of data involved in the training and validation of the models so further studies with larger datasets are needed to better implement machine learning approaches and AI technology.

1.2.2 Chen et al. (2022)

On the methodology suggested by Chen, Jian, Chi, *et al.* [13] the initial sample consisted of 789 male patients who underwent transrectal ultrasound-guided prostate

biopsy at the First Hospital of Jilin University between January 2013 and January 2021. A final total of 551 patients were included in the study. All patient data were collected through electronic medical records, including age, BMI, hypertension, diabetes, total PSA (tPSA), free PSA (fPSA), the ratio of serum fPSA to tPSA (f/tPSA), prostate volume (PV), PSA density (PSAD), neutrophil-to-lymphocyte ratio (NLR), and pathology reports of prostate biopsy. All examinations were completed within one week before the prostate biopsy.

Four supervised machine learning algorithms were used to build five PCa prediction models: tPSA univariate logistic regression (LR), multivariate LR, decision tree (DT), random forest (RF), and support vector machine (SVM). Three-quarters of the dataset was used for training, and the remaining observations served as the test set. The five prediction models were compared based on model performance metrics, such as the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, calibration curve, and clinical decision curve analysis (DCA).

The findings demonstrated that the RF, DT, and multivariate LR models showed better discrimination in the training set than the tPSA univariate LR and SVM models, with AUCs of 1.0, 0.922, and 0.91, respectively. Additionally, the multivariate LR model had the best discrimination in the test set (AUC=0.918). With no difference in performance between the training and test datasets, the multivariate LR model and SVM model exhibited greater extrapolation and generalizability. The other four models showed better net clinical benefits when compared to the DCA curves of the tPSA LR model.

1.2.3 Wang et al. (2022)

Regarding the study done by Wang [14], the Prostate Cancer dataset from the Population Health Data Archive (PHDA) was used, which included 1000 male patients with 427 prostate cancer cases and 573 prostatic hyperplasia cases. For modeling and prediction, different types of data were used: demographic information, prostate indicators, serum enzymes examinations, blood biochemical indicators and electrolyte indicators.

Random forest (RF), support vector machine (SVM), back propagation neural network (BP), and convolutional neural network (CNN) were used to predict the risk of PCa, among which BP and CNN were used on the enhanced data by SMOTE. The performances of models were compared using area under the curve (AUC) of the receiving operating characteristic curve. After the optimal model was selected, Shiny was used to develop an online calculator for PCa risk prediction based on predictive indicators.

Among the four models, RF had the best performance in predicting PCa (accuracy: 96.80%; AUC: 0.975, 95% CI: 0.964-0.986). Followed by BP (accuracy: 85.36%; AUC:

0.892, 95% CI: 0.849-0.934) and SVM (accuracy: 82.67%; AUC: 0.824, 95% CI: 0.805-0.844). CNN performed worse (accuracy: 72.37%; AUC: 0.724, 95% CI: 0.670-0.779).

1.2.4 Checcucci et al. (2021)

A total of 1447 patients were included in the dataset used in the study carried out by Checcucci, Rosati, De Cillis, *et al.* [15], of whom 623 were classified as "negative" (class 0, no risk of PCa), and 824 as "positive" (class 1, risk of PCa). The dataset includes the values for each patient's PSA, PSA density, previous prostate biopsies, number of suspicious lesions at mp-MRI, lesion volume, lesion location, and Pi-Rads score together with data from eight pre-bioptic factors.

A Fuzzy Interference System (FIS) was built with one output variable—the class—and eight input variables—corresponding to the eight pre-bioptic variables—that were each described by two triangular membership functions.

The results showed that 484 patients, or 33% of the total, were not classified, whereas 963 patients, or 67% of the total, were successfully classified. This ratio of unclassified patients is the biggest drawback of the study. Sensitivity (90.8%), specificity (59.2%), PPV (76.6%), NPV (81.3%), and AUC (0.77) were used to evaluate the performance of the classifier (for classified patients). Focusing on the most severe cases of prostate cancer, i.e., those with ISUP scores above 3, the model was also capable of accurately predicting the results of the biopsy in 98.1% of cases (accuracy).

1.2.5 Yu et al. (2021)

The study by Yu, Tao, Dong, *et al.* [16] included a sample of 688 patients with tPSA values less than 50 ng/ml who had never had a biopsy before. Of these patients, 480 (70%) were included in the Training set, while the remaining 208 patients (30%) were included in the Validation set. Age, PSA derivatives, prostate volume (PV), and mpMRI data were used as the patients' input variables. An analysis was conducted to see whether the available variables were positively or negatively influential.

The performance of four alternative machine learning techniques as Artificial Neural Network (ANN), Support Vector Machine (SVM), Classification And Regression Tree (CART), and Random Forest (RF) is evaluated in the study and compared with the results of Logistic Regression (LR) in the prediction of prostate cancer (PCa) and clinically significative prostate cancer (csPCa).

The various classifiers' diagnostic accuracy was assessed using the AUC calculation and Decision Curve Analysis (DCA). SVM, RF, and LR all achieved higher AUC values

for PCa prediction than ANN and CART (AUC = 0.891 and 0.834, respectively). SVM (AUC = 0.925), LR (AUC = 0.917), RF (AUC = 0.916), ANN (AUC = 0.911), and CART (AUC = 0.867) all show similar results when attempting to predict csPCa. However, it is remarkable to see that the CART technique outperforms all others and gives excellent concordance between projected and actual risk from the calibration plots for PCa and csPCa.

Finally, net-benefit curves were used to evaluate each model's clinical usefulness. From these curves, it can be shown that RF, SVM, ANN, and LR all exhibit similar tendencies and are, in any case, always more effective than CART, with probabilities between 0.05 and 0.4.

1.2.6 Takeuchi et al. (2019)

The aim of the prospective study by Takeuchi, Hattori-Kato, Okuno, *et al.* [17] was to compare the performance of Logistic Regression (LR) and Multilayer ANN in the prediction of prostate cancer. The dataset comprises 334 patients, each undergoing mpMRI, from which 232 patients (70%) were included in the Training set, validated by five-fold cross-validation, while the remaining 102 patients (30%) were included in the Test set.

All 22 original input variables (clinical variables) were used to train the classifier, whereas 12 were chosen using Lasso regression analysis and 9 were chosen using stepwise logistic regression analysis. The output variable makes a distinction between the presence of csPCa, any degree of PCa, and the lack of PCa. The Multilayer Artificial Neural Network (ANN) algorithm was trained for a range of hidden layer values (between 2 and 5, with each layer comprising 5 neurons) and step cycle values (1000, 2000 and 5000). The performance of the models was evaluated using the accuracy, the AUC the percentage of missing cancer, the Negative Predicted Values (NPV) and the Decision Curve Analysis (DCA).

The findings demonstrate that ANN performs best when 9 variables chosen by stepwise logistic regression, 5 hidden layers, and 2000 step cycles are used. The implemented model offers an accuracy value of 70.6%, which is 5–10% higher than the LR's, and an AUC value of 0.76, which is also higher than the LR's. The missed cancer rates are greater for LR (18% for any PCa and 9% for csPCa) than they are for ANN (16% for any PCa and 6% for csPCa). Next, the ANN displays NPV values of 76% (any PCa) and 94% (csPCa), as opposed to the LR's values of 72% (any PCa) and 91% (csPCa). In addition, the net-benefit curves built demonstrated that ANN is able to deliver greater net clinical benefit than LR.

2 THE DATASET

The dataset utilized for this project includes information gathered at the San Luigi hospital in Orbassano, which at the beginning had 1621 patients. These data were completely anonymous, with an ID code serving as the only way of identification.

In terms of the columns, there were 24 variables (including the first column, which was the patient ID) relevant to the investigation of prostate cancer:

- Age **(col. 2)**: age of the subject at the time of the study.
- TURP **(col. 3)**: Trans-Urethral Resection of the Prostate. Corresponds to a binary variable that determines whether a patient had a portion of the prostate surgically removed or not (1 indicates yes, 0 indicates no).
- PSA **(col. 4)**: Prostate-Specific Antigen [ng/ml].
- PSA density **(col. 5)**: calculated as the division between PSA and prostate volume, the units used were [ng/ml/cc].
- DRE **(col. 6)**: Digital Rectal Examination. This is a categorical variable resulting from the rectal examination intervention. A value of 0 indicates that the examination was negative, 1 that it was positive and therefore abnormalities were found, and 2 that the result is uncertain.
- DRE abnormality lobe **(col. 7)**: categorical variable that is taken into account only when the DRE was positive or uncertain, as it corresponds to the site where the abnormality is located. The possible values of the variable are: 0 for left lobe, 1 for right lobe, 2 for both lobes and finally, X if not applicable.
- Previous prostate biopsies **(col. 8)**: number of biopsies the patient has previously undergone.
- Result of I-II-III biopsy **(col. 9-11)**: categorical variable that reflects the result of the last 3 biopsies performed on the patient, if applicable. The values that can be taken are: 0 for normal prostate tissue, 1 for prostatitis, 2 for High-Grade Prostatic Intraepithelial Neoplasia (HGPIN), 3 for Atypical Small Acinar Proliferation (ASAP), 4 for tumour, and X if not applicable as the patient had not performed the biopsy.

- Number of suspicious lesions **(col. 12)**: number of lesions found.
- Prostate volume **(col. 13)**: prostate volume presented in [cc].
- I-II-III lesion volume **(col. 14-16)**: volume of the different lesions found presented in descending order according to their volume. That means that lesion I corresponds to the largest lesion, while lesion III refers to the smallest one, if applicable.
- Lesion location **(col. 17)**: area of the prostate in which the suspected lesion is located. The possible values that can be assumed by the variable are: 1 for the peripheral zone, 2 for the transitional zone (between the peripheral and the central zone), and 3 for the central zone.
- Lesion side **(col. 18)**: prostate side in which the suspicious lesion is located. Possible values are: 0 for the right side, 1 for the left side, and 2 if the lesion was located bilaterally.
- Lesion diameter **(col. 19)**: diameter of the largest lesion determined by MRI [mm].
- PIRADS **(col. 20):** score from 1 to 5 indicating the likelihood of clinically significant cancer.
- Gleason Score I (GS I, **col. 21**): Gleason grade of the most predominant pattern (can take values from 1 to 5).
- Gleason Score II (GS II, **col. 22**): Gleason grade to the second most predominant pattern (can take values from 1 to 5).
- Gleason Score total (GS tot, **col. 23**): sum of the two partial Gleason Scores (can take values from 2 to 10).
- ISUP (**col. 24**): system for grading prostate cancer between 1 and 5 depending on the Gleason Score.

The last 5 variables mentioned (col. 20-24) can assume the value of 'X' when it is not applicable, i.e. when the lesion is not associated with cancer.

2.1 Analysis of the dataset

Based on the dataset provided, it was decided to analyze the variables according to descriptive statistics. The dataset contained quite heterogeneous and different variables, so dividing the variables by type was necessary; based on this division, 3 different methods were adopted: estimation of the median, estimation of the mean, and computation of the frequency tables.

2.1.1 Computation of the median

This type of analysis was carried out with continuous integer and categorical ordinal variables, which correspond to the following:

- Age (continuous integer)
- Previous prostate biopsies (ordinal)
- Number of suspicious lesions (ordinal)
- Lesion diameter (continuous integer)
- PIRADS (ordinal)
- Gleason Score I (ordinal)
- Gleason Score II (ordinal)
- Gleason Score tot (ordinal)
- ISUP (ordinal)

The results for each variable are shown in the Table 2.1.

Table 2.1. Statistical description for continuous integer and ordinal input variables

Variable	Median
Age (col. 1)	70
Previous biopsies (col. 8)	0
Number of suspicious lesions (col. 12)	1
Lesion diameter (col. 19)	10
PIRADS (col. 20)	4
Gleason Score I (col. 21)	3
Gleason Score II (col. 22)	4
Gleason Score total (col. 23)	7
ISUP (col. 24)	2

2.1.2 Computation of the mean

In this section, the analysis was carried out based on the calculation of the mean and standard deviation, which applied only for real continuous variables, corresponding to those presented below:

- PSA
- PSA density
- Prostate volume

- I lesion volume
- II lesion volume
- III lesion volume

The values computed of these variables are represented in Table 2.2.

Table 2.2. Statistical description for continuous real input variables

Variable	Mean	Standard deviation
PSA (col. 4)	8.392	7.290
PSA density (col. 5)	0.174	0.170
Prostate volume (col. 13)	56.139	28.499
First lesion volume (col. 14)	0.968	1.839
Second lesion volume (col. 15)	0.381	0.413
Third lesion volume (col. 16)	0.762	1.060

2.1.3 Computation of the frequency tables

The third and final method consisted of analyzing the categorical and binary variables according to the frequency of each of their possible values. The variables that were taken into account are:

- TURP (binary)
- DRE (categorical non ordinal)
- DRE abnormality lobe (categorical non ordinal)
- Result of I biopsy (categorical non ordinal)
- Result of II biopsy (categorical non ordinal)
- Result of III biopsy (categorical non ordinal)
- Lesion location (categorical non ordinal)
- Lesion side (categorical non ordinal)

The frequency tables of every variable mentionated above are shown on the following Tables, from 2.3 to 2.10.

Table 2.4. Frequency table for the variable "DRE"

DRE (col. 6)	
Value	Frequency
"0"	1350
"1"	194
"2"	76
"NaN"	1

Table 2.5. Frequency table for the variable "DRE abnormality lobe"

DRE abnormality lobe (col. 7)	
Value	Frequency
"0"	103
"1"	151
"2"	10
"X"	1357

Table 2.3. Frequency table for the variable "TURP"

TURP (col. 3)	
Value	Frequency
"0"	1318
"1"	97
"NaN"	206

Table 2.6. Frequency table for the variable "result of the I biopsy"

Result of I biopsy (col. 9)	
Value	Frequency
"0"	398
"1"	12
"2"	35
"3"	40
"4"	193
"5"	0
"X"	1357

Table 2.7. Frequency table for the variable "result of the II biopsy"

Result of II biopsy (col. 10)	
Value	Frequency
"0"	158
"1"	4
"2"	9
"3"	10
"4"	14
"5"	0
"X"	1426

Table 2.8. Frequency table for the variable "result of the III biopsy"

Result of III biopsy (col. 11)	
Value	Frequency
"0"	39
"1"	0
"2"	3
"3"	4
"4"	2
"5"	0
"X"	1573

Table 2.9. Frequency table for the variable "lesion location"

Lesion location (col. 17)	
Value	Frequency
"0"	0
"1"	1282
"2"	311
"NaN"	0

Table 2.10. Frequency table for the variable "lesion side"

Lesion side (col. 18)	
Value	Frequency
"0"	842
"1"	768
"2"	0
"NaN"	11

2.2 Verification and modification of the dataset

After statistically analyzing the dataset given, it was necessary to make some verifications and modifications in order to have a corrected and reduced dataset, which would be used from now on. In order to compare the two datasets (the original and the modified one), the decision was taken to base the comparison on the missing values of each of the variables, because this number constitutes an important factor when applying a classifier to the data available. The results found for the first dataset are shown in Table 2.11.

Table 2.11. Number of missing values for each variable on the original dataset

Variable	Number of missing values
Age	0
TURP	206
PSA	5
PSA density	15
DRE	1
DRE abnormality lobe	1
Previous prostate biopsies	0
Result of I biopsy	0
Result of II biopsy	0
Result of III biopsy	0
Number of suspicious lesions	0
Prostate volume	9
I lesion volume	8
II lesion volume	0
III lesion volume	0
Lesion location	0
Lesion side	11
Lesion diameter	117
PIRADS	87
Gleason Score I	0
Gleason Score II	0
Gleason Score total	0
ISUP	0

As a first step, it was necessary to verify that the data for the variables I-II-III lesion volume (col. 14-16) were correct, which means that the order of these variables was checked, since, as mentioned in the explanation of the variables, the first one should correspond to the lesion with the largest volume. After performing this procedure, 40 patients were found to have the wrong order, so the order was corrected.

Subsequently, taking into consideration that PSA density can be calculated as the division between PSA and prostate volume, all those patients who had 2 of the 3 variables were taken and the missing variable was calculated. In this case, only 1 patient was found to be eligible for the above.

Finally, it was decided to merge the variables "Result of I-II-III biopsy" (col. 9-11) to

create a new variable called "Biopsies result" with the following specifications:

- 0: no biopsy done (to take the place of 'X').
- 1: at least one positive biopsy.
- 2: all biopsies negative.

After the modifications made to the dataset, the missing values for each variable were again calculated and are shown in Table 2.12.

Table 2.12. Number of missing values for each variable on the modified dataset

Variable	Number of missing values
Age	0
TURP	206
PSA	5
PSA density	14
DRE	1
DRE abnormality lobe	1
Previous prostate biopsies	0
Biopsies result	0
Number of suspicious lesions	0
Prostate volume	9
I lesion volume	8
II lesion volume	0
III lesion volume	0
Lesion location	0
Lesion side	11
Lesion diameter	117
PIRADS	87
Gleason Score I	0
Gleason Score II	0
Gleason Score total	0
ISUP	0

When comparing the amount of missing values before and after the modifications, it is evident that these are minimally reduced, and continue to have a high value, especially for the variables "TURP", "Lesion diameter" and "PIRADS". However, changes to the dataset are an important step before performing any procedures, because they can help ensure correct results later.

3 IMPUTATION OF THE VARIABLES

Missing values can result from data loss, participant dropouts, and nonresponses, among other things. Missing values result in a smaller sample size than expected, which eventually compromises the validity of the study's findings. When conclusions about a population are made based on such a sample, it can also lead to skewed results, weakening the validity of the data. Elimination and imputation are typically applied when missing values are present in a dataset, but deleting data is not recommended because it can lead to a loss of important parameters; instead, to create a more complete dataset, imputation entails substituting values using different methods such as statistic, regressions or predictors [18].

Due to the high number of missing values found in the dataset provided for this project, before applying a classifier, it was decided to implement a variable imputation using the k-Nearest Neighbors (kNN) algorithm, which is a machine learning technique based on supervised learning. Basically, the working principle of this technique consists of calculating the distance between an observation and the rest of the dataset, taking the "k" closest elements, and from the majority voting, determining which class it belongs to. The metric to calculate the distance between the data must be chosen appropriately, evaluating the type of variables available. In addition, the configuration of the parameter k is crucial: if k is chosen too small, the class is assigned on the basis of the few neighboring elements, and thus the classification may be influenced by data noise; if, on the other hand, k is chosen too big, the element will be classified in the most numerous class in the entire population, thus underestimating the concept of proximity between the data [19].

In this case the patients were divided into 6 different groups according to the specific value of the ISUP before implementing the imputation, in order to only delivering to the classifier values associated with the patient's condition, i.e. for a patient with an ISUP of 2, values of variables in accordance with his condition would be predicted. These groups were:

- Group 1: ISUP equal to 1

- Group 2: ISUP equal to 2
- Group 3: ISUP equal to 3
- Group 4: ISUP equal to 4
- Group 5: ISUP equal to 5
- Group 6: ISUP equal to 'X' (non applicable)

Furthermore, from each of the groups it was decided to calculate a new subgroup, which was necessary for the imputation, and that contained only those patients who did not have any missing values. These generated groups had different numbers of components, and it was determined to make two classifiers for each group of subjects, one with k=5% of the number of patients per group, and the other with k=10%. The imputations made were organized in the following order:

- **Imputation 1:** k=5 % of the total number of patients in each group.
- **Imputation 2:** k=5 % of the reduced groups (only patients without missing values in each group).
- **Imputation 3:** k=10 % of the total number of patients in each group.
- **Imputation 4:** k=10 % of the reduced groups (only patients without missing values in each group).

Detailed information of the number of people per group and the respective value of k is presented in Table 3.1.

Subsequently, the kNN for the imputation was used for the variables that showed off missing values, which are shown in Table 2.12, so that a different model was generated for each of them, thus generating 10 different models per k value, for a total of 40 models.

Table 3.1. Description of the different models used for the imputation of variables

Value of ISUP	Number of subjects		k=5%		k=10%	
	Total	Without missing	Total	Without missing	Total	Without missing
1	120	94	6	5	12	9
2	420	346	21	17	42	35
3	214	185	11	9	22	18
4	109	97	5	4	10	9
5	50	46	3	2	6	4
X	708	574	35	28	70	57

3.1 Boxplot

In order to evaluate the imputations previously carried out, boxplots were made for each variable, in which the dispersion of only the imputed data, only the original data, and finally, the original data together with the imputed data for each of the 6 groups created from the ISUP were shown in the same figure. All these graphs are reported in the *appendix*.

The most important thing found at this point was that of the 10 variables that contained missing values, only 6 could be boxplotted, as the others had a few fields to be imputed . The information on the excluded variables and those that were taken into account are shown in Table 3.2.

Table 3.2. Variables taken into account for the boxplot

Variables boxplotted	Variables excluded
-Lesion diameter	
-Lesion side	-DRE
-PIRADS	-DRE abnormality lobe
-Prostate volume	-First lesion volume
-PSA Density	-PSA
-TURP	

Additionally, through this graphical representation, it could be corroborated that the imputed values did not strongly influence the statistical characteristics of the original groups and what they generated in some cases was a displacement of the median, which is a positive factor when using the dataset as input to a classifier.

3.2 Exclusion of variables and creation of new datasets

Based on the results of the imputation and the behaviour of the graphs presented, the decision was taken to definitively exclude from the imputation some variables that were divided into 2 types:

1. Variables that had less than 5 imputed values, which corresponded to:

    - DRE
    - DRE abnormality lobe
    - PSA

2. Clinical variables that are determined by a doctor on the basis of pre-established criteria:

    - Lesion side
    - PIRADS
    - TURP

In the case of the first type of variables, the patients with missing values were directly eliminated, which led to the removal of 7 rows from the dataset, while when taking into account the variables belonging to the second type, it was decided to opt for two different paths (which led to the creation of two different datasets: "Dataset 1" and "Dataset 2") in order to subsequently see how the behaviour of the classifier would change. In "Dataset 1" what was done was to eliminate exactly the 3 variables in consideration, and in "Dataset 2" an analogous procedure to the one done with the first type of variables was carried out, i.e. only eliminating the patients who had missing values in these variables instead of removing the whole variables, leading to the suppression of 249 subjects from the dataset.

After carrying out these procedures, Dataset 1 was left with 1614 patients and 19 variables, while Dataset 2 was finally established with 1365 subjects with 22 variables available.

4 CLASSIFICATION

Prior to the application of the classifier, the outcome variable or 'class' was created by assigning a value of 0 (negative class) to all subjects who had a negative prostate biopsy result and are considered cancer-free subjects (all those for whom the ISUP measurement did not apply, i.e. who had a value of 'X'), while all subjects who had a positive biopsy were assigned a class value of 1 (ISUP with a numerical value, between 1 and 5). The datasets were distributed as follows:

- : **Dataset 1** : 705 subjects in class 0 and 909 in class 1.
- : **Dataset 2** : 584 subjects in class 0 and 781 in class 1.

It can be seen that the data are slightly unbalanced because class 1 has a higher representation (approximately 60% of all patients) but this does not really lead to a significant compromise in classifier performance, which should not cause a class bias.

4.1 kNN description

Having already defined the outcome variable, the kNN was implemented again, this time as a classification method. This algorithm is normally trained using a subset of the dataset, called the training set, which contains all the relevant variables including the class to which each of the patients present belongs, and then, from this, a test is performed with the observations that were not part of the training set, finally giving an estimate of the classifier's ability to predict the outcome of the unlabelled data.

Besides, as previously anticipated, the decision of which distance to use when implementing the classifier is a critical step that can have a great influence on its performance. As far as our dataset is concerned, it is evident that we are in the presence of very heterogeneous variables, which fall into different types. The use of a standard distance (or similarity) measure, such as the Manhattan, the Euclidean or the Chebyshev, is not recommended in these circumstances, while the use of different distances that take into account other specific parameters can be effective. More specifically, from intermediate

tests, it was concluded that it was convenient to use 2 distances and evaluate how the results could change. These distances will be explained below.

4.1.1 Gower distance

The Gower distance was developed in 1971, especially for use when in possession of mixed data, i.e. a dataset that may contain a combination of numerical, categorical or binary information. This similarity index substantially unifies Jaccard's coefficient which is used for binary variables, the simple matching coefficient used for multistate categorical variables and normalized city block distance used for quantitative variables.

More specifically, given two p-dimensional vectors z_i and z_j, Gower's similarity coefficient is defined as:

$$s_{ij} = \frac{\sum_{h=1}^{p_1}(1 - \left|z_{ih} - z_{jh}\right| / R_h) + a + \alpha}{p_1 + (p_2 - d) + p_3}, 0 \leq s_{ij} \leq 1 \tag{4.1}$$

where $p = p_1 + p_2 + p_3$ is the total number of variables, p_1 is the number of continuous variables, a and d are the number of positive and negative matches, respectively, for the p_2 binary variables, α is the number of matches for the p_3 multi-state categorical variables, and R_h is the range of the h-th continuous variable [20].

As can be seen from the above expression, for continuous variables, the size of the gap between two values assumed by the variable has a bearing on the calculation of the overall distance, whereas for categorical and binary variables it matters whether the two values are a match or not.

In this case, as MATLAB was used as the software to carry out the calculations and as it does not have Gower's distance among those that are included in the tool for the function *'fitcknn'*, a new algorithm had to be created. Basically, it calculates the Gower distance between a vector X and a matrix Y, taking into account the type of variable of each input. These input arguments could be of matrix type or table type. For elements delivered in table form, a vector had to be added to determine the type of variable: 1 for quantitative, 2 for binary and 3 for categorical, while, if they were delivered in array form, the algorithm was able to predict the type of variable of each of the elements, so the aforementioned vector was not necessary. The creation of this function was based on a similar algorithm present in the MATLAB FSDA Toolbox [21].

4.1.2 Hamming distance

Hamming distance is a distance metric that measures the number of mismatches between two vectors. It is mostly used for nominal data, string and bit-wise analyses, and also can be useful for numerical data [22]. This is the default distance used by MATLAB when given a table as input argument, containing several types of variables and calculated as :

$$Hamming(x,y) = \sum_{i=1}^{n} 1_{x_i \neq y_i} = \sum_{i=1}^{n} |x_i - y_i| \tag{4.2}$$

Analyzing this expression, it can be noted that the Hamming distance method looks at the whole data and finds when data points are similar and dissimilar one to one, and at the end it gives the result of how many attributes were different.

4.2 kNN implementation

As previously introduced, 2 different global datasets were taken, containing the 4 specific datasets for each of the imputations performed, to which the kNN algorithm with the 2 chosen distances (Hamming and Gower) would be applied. It was decided to select a third dataset, whose results would be used for purely comparative purposes, in which only the subjects who did not present any missing value in any variable were taken.

In addition to this, the classifier was implemented with the data in two different modes: taking them as originally reported and normalising them using min-max scaling. The normalisation was selected because it was necessary to avoid that the size of the ranges influenced the calculation of the distance between two elements, resulting in an incorrect classification of the elements. For the min-max scaling, the following equation is used:

$$x_{norm} = \frac{x - min(x)}{max(x) - min(x)} \tag{4.3}$$

Where x corresponds to the vector containing all the data of a given variable. This means that the operation had to be repeated i times, where i is the number of variables to be taken into account as input for the classifier.

It is important to highlight that some of the available observations directly determined the presence of cancer because they took the value of 'X' (non applicable) when there was no cancer, while if any numerical value was taken, this indicated that the sub-

ject under consideration had a tumor. As a result, these variables could not be used as input to the classifier in the case of our interest. These variables were removed from each dataset and corresponded to:

- Gleason Score I
- Gleason Score II
- Gleason Score total
- ISUP

As mentioned, this decision was made precisely because the objective of the classifier was to directly predict whether the person had cancer or not, however, these removed variables can be very useful if we want to evaluate other conditions or other types of output, such as clinically significant cancer, where a multiclass classifier could be used.

In terms of model validation, k-fold cross validation was selected, which consists of dividing the dataset into k different subgroups of approximately the same dimension and then applying the classifier: folds are used for model construction and the hold-out fold is allocated to model validation [23]. In this case, the dataset was divided into 10 subset.

On the other hand, speaking about the value of k chosen for the kNN, it was originally decided to take it as suggested in the literature $k = \sqrt{N}$ to the nearest odd number, where N is the number of subjects present in the dataset. However, 3 other values of k, corresponding to 75 %, 50 % and 25 % of the initial value of k, were subsequently tested to evaluate if there were changes on the performance. All of the above led to the training and application of 144 different classifiers, varying the datasets, imputations, distances and k values.

Finally, in order to adequately evaluate and compare the performance of the various models, the confusion matrix was defined for each classification made, obtained by comparing the predicted output with the original output and calculating the number of true positives (values classified in class 1 and belonging to class 1), true negatives (values classified in class 0 and belonging to class 0), false positives (values classified in class 1 and belonging to class 0) and false negatives (values classified in class 0 and belonging to class 1). Based on these values, a certain number of descriptive parameters were then calculated, which were useful for the interpretation of the results. These parameters are [24]:

- **Specificity:** percentage of items correctly classified as negative out of the total of truly negative items.

$$specificity = \frac{TN}{TN + FP} * 100\% \tag{4.4}$$

- **Sensitivity:** percentage of items correctly classified as positive out of the total of truly positive items.

$$sensitivity = \frac{TP}{TP + FN} * 100\% \tag{4.5}$$

- **Positive Predicted Value (PPV):** percentage of items correctly classified as positive out of the total of items classified as positive.

$$PPV = \frac{TP}{TP + FP} * 100\% \tag{4.6}$$

- **Negative Predicted Value (NPV):** percentage of items correctly classified as negative out of the total of items classified as negative.

$$NPV = \frac{TN}{TN + FN} * 100\% \tag{4.7}$$

- **Balanced accuracy:** is a metric used to evaluate how well classification models perform on imbalanced datasets. It guarantees that both minority and majority classes are given equal weight when being evaluated because it represents the arithmetic mean of sensitivity and specificity [25].

$$Balanced\ accuracy = \frac{1}{2}(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}) * 100\% \tag{4.8}$$

$$Balanced\ accuracy = \frac{1}{2}(sensitivity + specificity) * 100\% \tag{4.9}$$

5 Results and analysis

The *appendix* reports the confusion matrix of the implemented classifiers, divided according to the global dataset used ("Dataset 1" and "Dataset 2"). It should be clarified that these figures show a class denoted as 1 that actually corresponds to the negative classification (absence of cancer) and another class expressed as 2 that is actually the positive class (presence of cancer).From now on and to avoid confusions, there will be reference only to class positive and negative, not with numbers.

To better demonstrate the results obtained and to be able to make a more effective comparison, some graphs of the specificity, sensitivity, NPV, PPV and balanced accuracy behavior were made, which can be seen in Figures 5.1, 5.2, 5.3, 5.4 and 5.5, respectively.

In the case of specificity (Figure 5.1), it should be highlighted that practically all classifiers, with the exception of 3 of the 4 classifiers constructed using the Gower distance, attained values above 55%. This low specificity value suggests that the classifiers had a difficult time identifying negative patients, which suggests that they frequently made positive predictions about subjects that had been provided to them as input; this can also be corroborated by looking at the confusion matrix shown in the appendix. In addition to this, analyzing the behavior of these 3 classifiers through the different imputations and k values, it can be seen that practically this constitutes a straight line, which means that the modifications made did not present any minimally significant change, resulting in almost the same number of patients in the different classes and a quite similar classification error.

Now when discussing sensitivity (Figure 5.2), it is evident that it behaves in a manner that is practically the opposite of that which was previously described when it comes to specificity, supporting the claims made. In this instance, we can see that the three classifiers mentioned had a fairly high sensitivity (between 65% and 92%), standing out among those performed with Datasets 1 and 2 without normalization, indicating that for these predictors, normalization is crucial [26] because they were unable to generalize the classification and determined that almost all subjects were positive, which is not useful for our objective because it would be precisely in line with the methods currently used

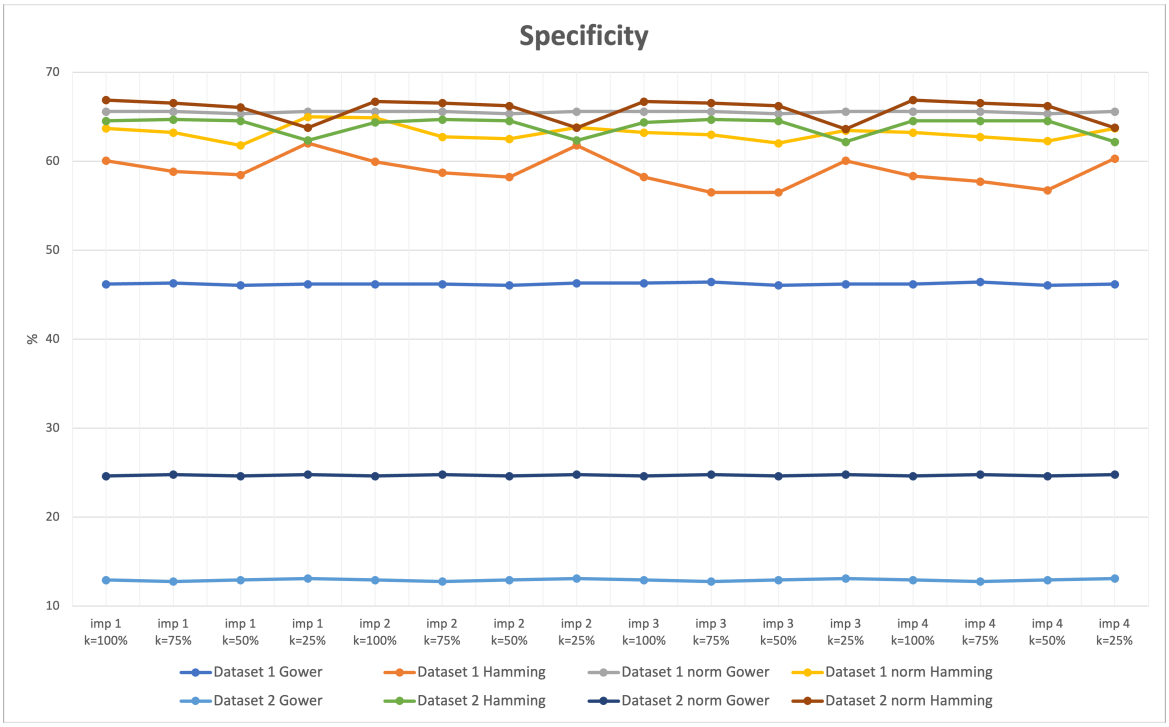


Figure 5.1. Specificity values calculated for each of the classifiers.

where many unnecessary biopsies are performed. On the other hand, it is noteworthy that Dataset 2 performed well when using the Hamming distance, both normalized and unnormalized, as they demonstrated an important level of specificity and this is paired with a high value of sensitivity, in addition to emphasizing that the change between the various imputations and the values of k chosen differences are found.

Here it could be seen that the classifier performed with the Gower distance that was not mentioned yet, that is, the normalized Dataset 1, presented the lowest sensitivity of all, being below 50 %, which is corroborated by appreciating its corresponding confusion matrix, where, for all the imputations and all the values of k, more preference was given to the negative class, classifying 966 patients out of 1611 in this class (approximately 60 % of the whole dataset) and thus being the only classifier with Gower's distance that behaved in the opposite way to the others.

On the other hand, referring to the NPV values (Figure 5.3), it was found that in general the classifiers performed with the Hamming distance for both datasets had a higher performance in this area, taking values ranging from approximately 60 % to 68 %, which also applies to the PPV results (Figure 5.4), with the highest values being those predictors that used Dataset 2 with Hamming distance, and those that were considerably farther from the ranges presented by the other classifiers.

Finally, speaking of the balanced accuracy (Figure 5.5), a parameter that refers to the general performance of the classifiers and that allows to express the accuracy presented

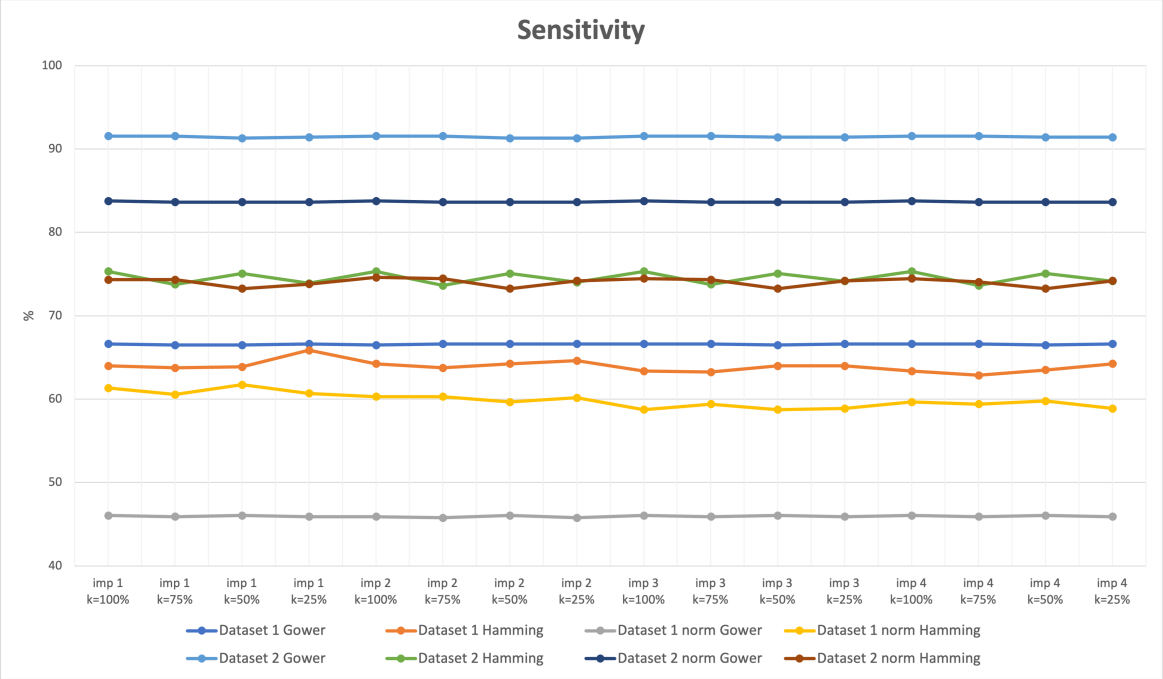


Figure 5.2. Sensitivity values calculated for each of the classifiers.

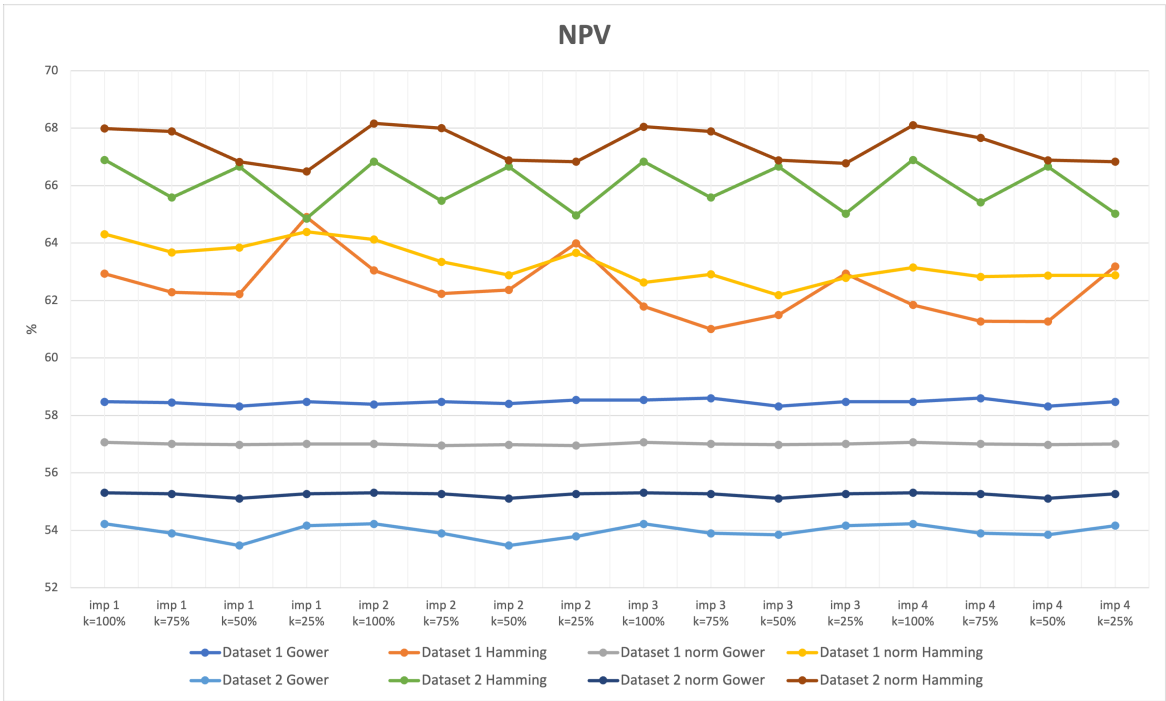


Figure 5.3. NPV values calculated for each of the classifiers.

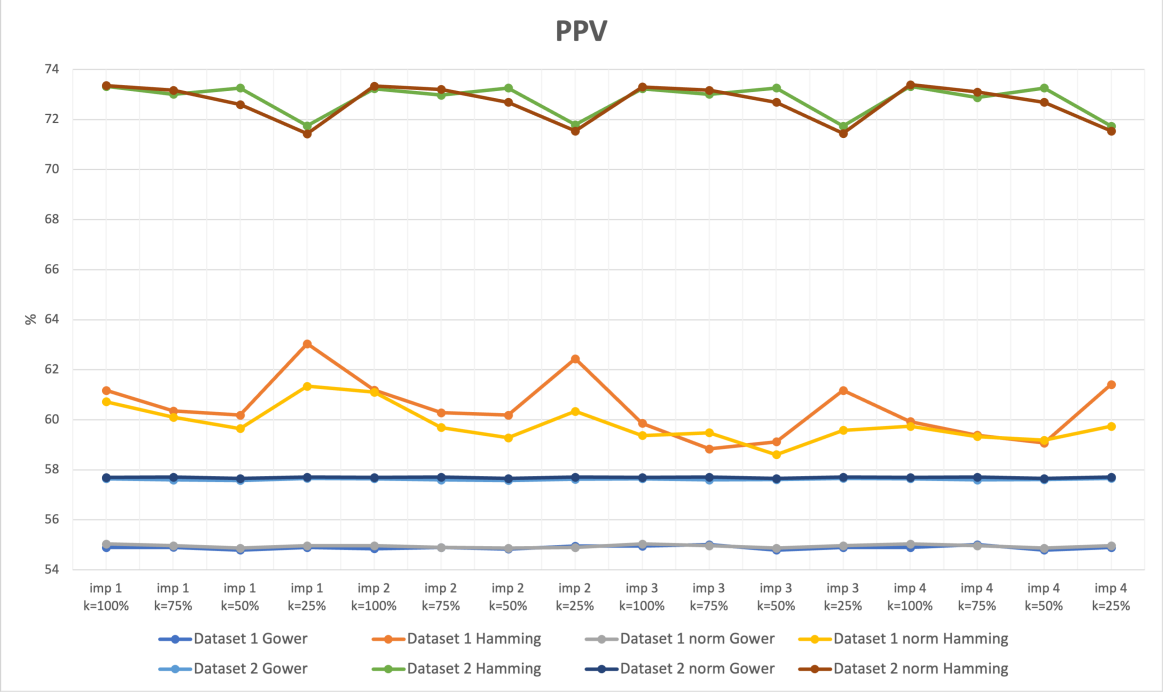


Figure 5.4. PPV values calculated for each of the classifiers .

in identifying both classes, it is found that again the classifiers that used Dataset 2 with Hamming distance presented the highest values (between 67% and 72%), while those with the lowest performance corresponded to those of Dataset 2 and Gower distance (between 52% and 54%).

As was already said, the main objective of this thesis project and the development of the classifiers is to achieve a completely non-invasive diagnosis based solely on clinical characteristics while simultaneously reducing the number of unnecessary biopsies. Because of this, it was decided to give the accurate classification of negative subjects more importance in order to prevent positive subjects who did have the pathology from being classified as negative, which would have a negative impact on the patient's quality of life and possibly jeopardize their survival. NPV and balanced accuracy are the two metrics that, based on these assumptions, may provide more insight into the classifiers: the value of the former is inversely proportional to the number of FNs, and thus, by preferring models with high NPV, it is possible to minimize the percentage of FNs, while the latter gives an average indication of the model's ability to correctly classify items belonging to both classes, considering that the dataset in our possession is not perfectly balanced, but has a greater representation of the positive class.

Taking into account the considerations previously made, it is clearly observed that there are 2 classifiers that stood out over the others, which corresponded to those performed with Hamming distance using Dataset 2, which consisted in the elimination of

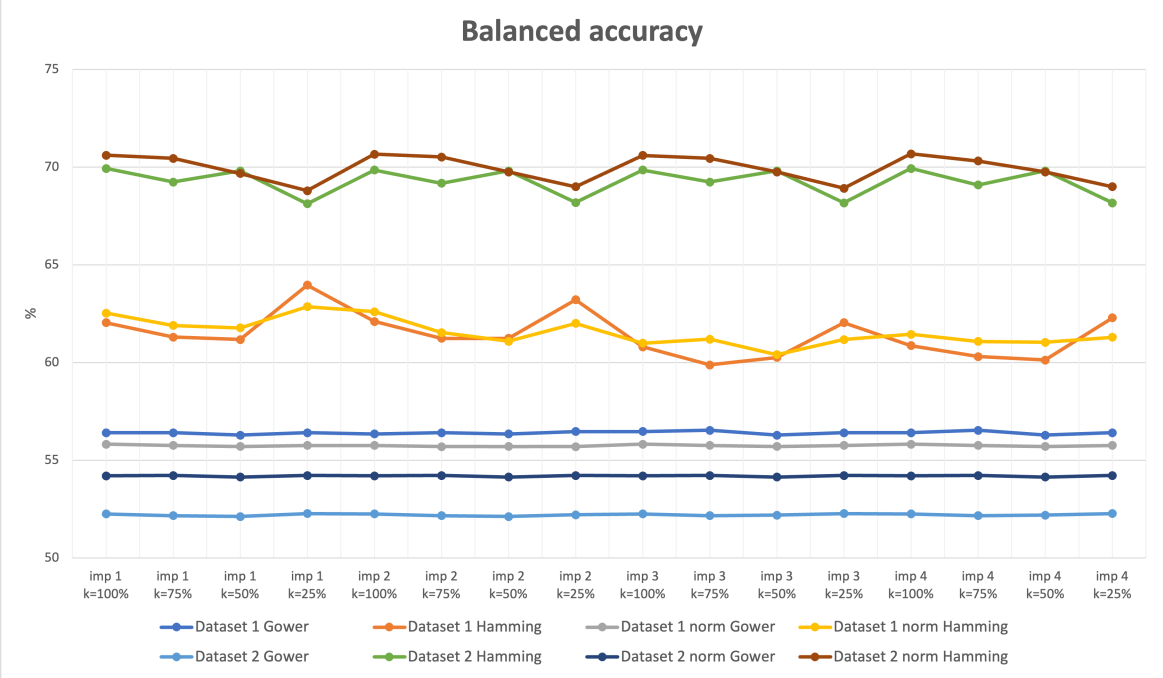


Figure 5.5. Balanced accuracy values calculated for each of the classifiers .

patients with missing values in the variables "Lesion side", "PIRADS" and "TURP", which shows that for this type of distance, the permanence of these variables in the dataset was of vital importance. On the other hand, it is evident that, in contrast, if the Gower distance is taken, the performance improves slightly using Dataset 1, in which the previously mentioned variables were directly eliminated.

Now that the classifiers with the best performance have been chosen in general, it is necessary to analyze which imputation showed the best results and for which k value. If we take into account that we are giving more importance to the performance evaluated by NPV and balanced accuracy, we can see that the predictor represented by the brown line (Dataset 2 normalized with Hamming distance) presented slightly higher values, the values found for all its modifications are presented in Table 5.1, where the 3 best performances that entail quite similar results are presented in bold, this implies that in reality the different imputations that were carried out were quite similar, which was corroborated by comparing the predicted values, where it was found that the highest percentage of discrepancies between the groups was less than 3%. Additionally, when analyzing these classifiers that presented the best results, it was found that the value of k corresponding to that reported in the literature (square root of N, which was the initial value) is the most appropriate to use in the classifier.

Finally, taking into account the last dataset employed, Figure 5.6 and Figure 5.7 report the confusion matrixes that corresponded to the results of the classifiers made for Dataset 3 which represented purely patients with no missing values, hence no impu-

Table 5.1. Performance of the global best classifier between Datset 1 and Dataset 2: the one made with Dataset 2 normalized and Hamming distance. Values are reported in percentage (%).

Classifier		Specificity	Sensitivity	NPV	PPV	Balanced accuracy
Imputation 1	k=100%	**66,88**	**74,34**	**67,99**	**73,36**	**70,61**
	k=75%	66,56	74,34	67,89	73,17	70,45
	k=50%	66,07	73,27	66,83	72,60	69,67
	k=25%	63,78	73,80	66,50	71,43	68,79
Imputation 2	k=100%	**66,72**	**74,60**	**68,17**	**73,33**	**70,66**
	k=75%	66,56	74,47	68,00	73,20	70,51
	k=50%	66,23	73,27	66,89	72,69	69,75
	k=25%	63,78	74,20	66,84	71,54	68,99
Imputation 3	k=100%	**66,72**	**74,47**	**68,05**	**73,30**	**70,59**
	k=75%	66,56	74,34	67,89	73,17	70,45
	k=50%	66,23	73,27	66,89	72,69	69,75
	k=25%	63,62	74,20	66,78	71,45	68,91
Imputation 4	k=100%	**66,88**	**74,47**	**68,11**	**73,39**	**70,68**
	k=75%	66,56	74,07	67,66	73,10	70,31
	k=50%	66,23	73,27	66,89	72,69	69,75
	k=25%	63,78	74,20	66,84	71,54	68,99

tations were made. In addition, Table 5.2 shows the performance results in terms of specificity, sensitivity, NPV, PPV and balanced accuracy. These results reaffirm that the Gower distance did not perform well as it practically classified most of the patients as positive for cancer, and finally that Hamming distance with the normalized dataset was shown to perform the best.

When talking about the preferences chosen for the classifier (high values of balanced accuracy and NPV), it is evident that "Dataset 3" presented the highest percentages, especially in NPV, which indicates that, in this case, to obtain better results it was enough to eliminate the patients with missing values and keep all the variables. Finally, this ratified that the most convenient k for the classifier corresponded to the initial one, i.e. the square root of the number of patients.

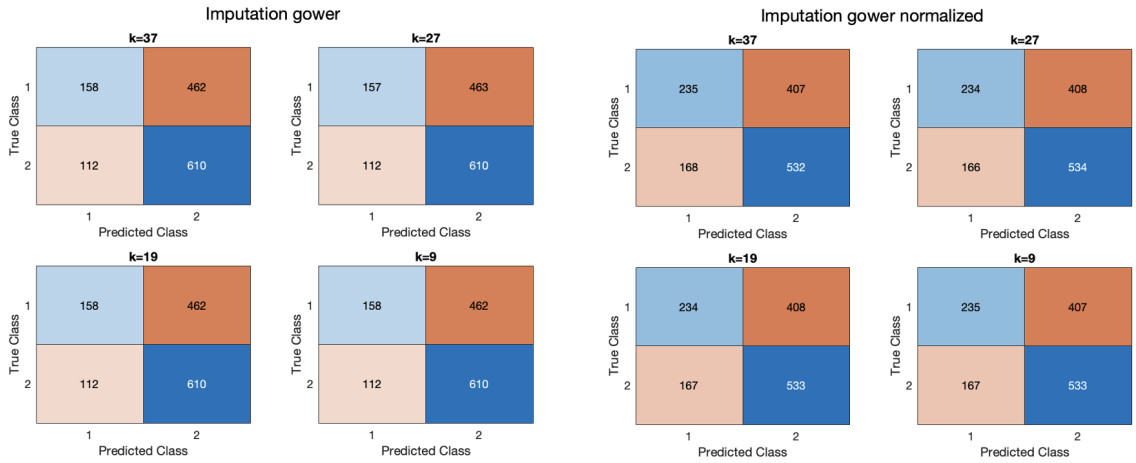


Figure 5.6. Confusion matrixes of the classifiers done with Gower distance, original and normalized, respectively.

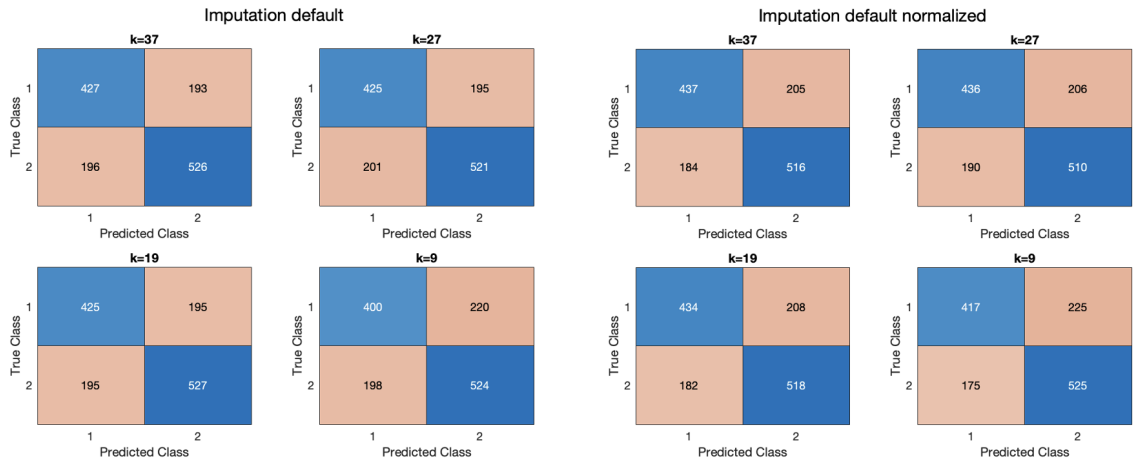


Figure 5.7. Confusion matrixes of the classifiers done with Hamming distance, original and normalized, respectively.

Table 5.2. Performance of the classifiers implemented with Dataset 3. Values are reported in percentage (%).

Classifier		Specificity	Sensitivity	NPV	PPV	Balanced accuracy
Gower	k=100%	25.48	84.49	58.52	56.90	54.99
	k=75%	25.32	84.49	58.36	56.85	54.91
	k=50%	25.48	84.49	58.52	56.90	54.99
	k=25%	25.48	84.49	58.52	56.90	54.99
Gower normalized	k=100%	36.60	76.00	58.31	56.66	56.30
	k=75%	36.45	76.29	58.50	56.69	56.37
	k=50%	36.45	76.14	58.35	56.64	56.30
	k=25%	36.60	76.14	58.46	56.70	56.37
Hamming	k=100%	68.87	72.85	68.54	73.16	70.86
	k=75%	68.55	72.16	67.89	72.77	70.35
	k=50%	68.55	72.99	68.55	72.99	70.77
	k=25%	64.52	72.58	66.89	70.43	68.55
Hamming normalized	k=100%	**68.07**	**73.71**	**70.37**	**71.57**	**70.89**
	k=75%	67.91	72.86	69.65	71.23	70.38
	k=50%	67.60	74.00	70.45	71.35	70.80
	k=25%	64.95	75.00	70.44	70.00	69.98

6 Conclusions

When it comes to prostate cancer, the developed project is positioned as a non-invasive diagnostic solution that can help in the identification of malignant tumours without exposing healthy patients to the discomfort and possible side effects of a biopsy.

Initially, the identification of the different types of variables within the dataset was helpful in highlighting the heterogeneity of the observations present, which was useful in deciding which of the possible distances to choose for the kNN algorithm. Later, in the development of this algorithm, the fact of making changes and verifications in the variables allowed the data that were then delivered to the different classifiers to be complete and without mistakes, so that better results were achieved in terms of performance. Then, the analysis of the missing values of each of the variables allowed the identification of a major problem in the dataset in possession, which effectively corresponded to the presence of several fields denoted as NaN.

In the second part of the project, attention was focused on how to handle these missing values and the decision was taken to perform an imputation through a neighbourhood-based algorithm (kNN), in order to generate a much more complete dataset. In this step it was decided to take different values of k to see what the influence on the final predictor would be. Here it was of crucial importance to separate the patients according to their ISUP value, as this would allow the imputation of the variables to be more in line with the actual state of the patient. After imputation, the statistical analysis of the generated values was of great help as it allowed to observe whether the created data changed the characteristics of the different groups, which resulted in parameters such as median, mean and quartiles being minimally affected or not modified. In the final stage of this part, the variables that did not make medical sense to impute were identified as "lesion side", "PIRADS" and "TURP", and 3 global datasets were generated, which differed in the way the variables were handled: eliminating the 3 variables (Dataset 1), eliminating the patients with missing values in these variables (Dataset 2) and keeping only the patients who in general had no missing values (Dataset 3).

Regarding the last part of the project, after defining the outcome variable, 2 differ-

ent distances were used for the kNN predictor (Hamming distance and Gower distance), both normalised and unnormalised datasets were taken and, additionally, 4 different values of k were taken into account for the neighbourhood, thus allowing a much broader analysis of the classifier. Looking at the results delivered after implementing the classifiers with the indicated parameter changes (144 classifiers in total) and evaluating their performance mainly in terms of NPV and balanced accuracy, it became evident that the Hamming distance using the normalised dataset delivered the best results, mainly when using datasets 2 and 3, showing that the best value of k corresponded to the initial value (square root of the number of patients) and showing that in fact the different imputations did not have a major effect on the classifier.

In conclusion, this research has shown how challenging it is to create a comprehensive decision-aid system that can accurately identify, among a large population of men at risk, those who actually have prostate cancer and thus require a biopsy.The findings were fairly encouraging: even though the classifiers' performance was not ideal, this research would be useful in the future for the development of better prostate cancer diagnostic tools, particularly when a dataset as heterogeneous as the one utilized in this project is present.

BIBLIOGRAPHY

[1] https://www.cancer.org/cancer/types/prostate-cancer.html., Accessed: 2023-9-02.

[2] *Prostate*, en, https://www.brandywineuc.com/practice-areas/general-urology-care/prostate/., Accessed: 2023-9-02, Feb. 2021.

[3] S. H. Selman, "The mcneal prostate: A review," *Urology*, vol. 78, no. 6, pp. 1224–1228, 2011, ISSN: 0090-4295. DOI: https://doi.org/10.1016/j.urology.2011.07.1395. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0090429511021376.

[4] R. J. Rebello, C. Oing, K. E. Knudsen, *et al.*, "Prostate cancer," en, *Nat. Rev. Dis. Primers*, vol. 7, no. 1, p. 9, Feb. 2021.

[5] *Prostate disease*, en, https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/prostate-disease, Accessed: 2023-9-02.

[6] O. Nettey, A. Walker, M. K. Keeter, *et al.*, "Mp21-17 black race predicts significant prostate cancer independent of clinical setting, and clinical and socioeconomic risk factors: Evidence for a biological basis of disparities," en, *J. Urol.*, vol. 199, no. 4S, e269, Apr. 2018.

[7] J.-L. Descotes, "Diagnosis of prostate cancer," en, *Asian J. Urol.*, vol. 6, no. 2, pp. 129–136, Apr. 2019.

[8] *Fondazione AIRC per la ricerca sul cancro ETS*, it, https://www.airc.it/cancro/informazioni-tumori/guida-ai-tumori/tumore-della-prostata., Accessed: 2023-9-12.

[9] *Cancer today*, en, https://gco.iarc.fr/today/home, Accessed: 2023-9-14.

[10] P. Rawla, "Epidemiology of prostate cancer," en, *World J. Oncol.*, vol. 10, no. 2, pp. 63–89, Apr. 2019.

[11] D. A. Roffman, G. R. Hart, M. S. Leapman, *et al.*, "Development and validation of a multiparameterized artificial neural network for prostate cancer risk prediction and stratification," en, *JCO Clin. Cancer Inform.*, vol. 2, no. 2, pp. 1–10, Dec. 2018.

[12] M. Gravina, L. Spirito, G. Celentano, *et al.*, "Machine learning and clinical-radiological characteristics for the classification of prostate cancer in pi-rads 3 lesions," *Diagnostics*, vol. 12, no. 7, 2022, ISSN: 2075-4418. DOI: 10.3390/diagnostics12071565. [Online]. Available: https://www.mdpi.com/2075-4418/12/7/1565.

[13] S. Chen, T. Jian, C. Chi, *et al.*, "Machine learning-based models enhance the prediction of prostate cancer," en, *Front. Oncol.*, vol. 12, p. 941 349, Jul. 2022.

[14] C. Wang, "Prostate cancer risk prediction and online calculation based on machine learning algorithm." chinese medical sciences journal = chung-kuo i hsueh k'o hsueh tsa chih," vol. 37, pp. 210–217, 2022.

[15] E. Checcucci, S. Rosati, S. De Cillis, *et al.*, "Artificial intelligence for target prostate biopsy outcomes prediction the potential application of fuzzy logic," en, *Prostate Cancer Prostatic Dis.*, vol. 25, no. 2, pp. 359–362, Feb. 2022.

[16] S. Yu, J. Tao, B. Dong, *et al.*, "Development and head-to-head comparison of machine-learning models to identify patients requiring prostate biopsy," en, *BMC Urol.*, vol. 21, no. 1, p. 80, May 2021.

[17] T. Takeuchi, M. Hattori-Kato, Y. Okuno, S. Iwai, and K. Mikami, "Prediction of prostate cancer by deep learning with multilayer artificial neural network," en, *Can. Urol. Assoc. J.*, vol. 13, no. 5, E145–E150, May 2019.

[18] S. K. Kwak and J. H. Kim, "Statistical data preparation: Management of missing values and outliers," en, *Korean J. Anesthesiol.*, vol. 70, no. 4, pp. 407–411, Aug. 2017.

[19] H. A. Abu Alfeilat, A. B. A. Hassanat, O. Lasassmeh, *et al.*, "Effects of distance measure choice on k-nearest neighbor classifier performance: A review," en, *Big Data*, vol. 7, no. 4, pp. 221–248, Dec. 2019.

[20] J. C. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, vol. 27, no. 4, p. 857, Dec. 1971.

[21] M. Riani, D. Perrotta, and F. Torti, "Fsda: A matlab toolbox for robust analysis and interactive data exploration," *Chemometrics and Intelligent Laboratory Systems*, vol. 116, pp. 17–32, 2012, ISSN: 0169-7439. DOI: https://doi.org/10.1016/j.chemolab.2012.03.017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169743912000974.

[22] R. W. Hamming, "Error detecting and error correcting codes," *The Bell System Technical Journal*, vol. 29, no. 2, pp. 147–160, 1950. DOI: 10.1002/j.1538-7305.1950.tb00463.x.

[23] Y. Jung, "Multiple predictingk-fold cross-validation for model selection," en, *J. Nonparametr. Stat.*, vol. 30, no. 1, pp. 197–215, Jan. 2018.

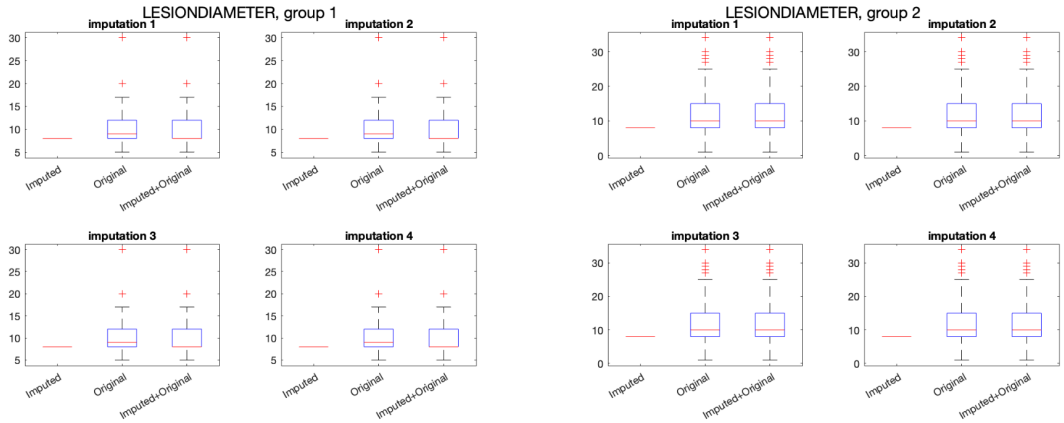
[24] E. Sharp, *Statistics*, en, https://geekymedics.com/sensitivity-specificity-ppv-and-npv/, Accessed: 2023-9-14, Jun. 2018.

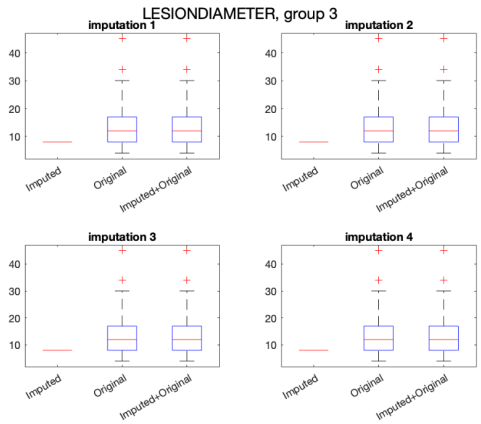
[25] *What is balanced accuracy?* en, https://www.educative.io/answers/what-is-balanced-accuracy, Accessed: 2023-9-14.

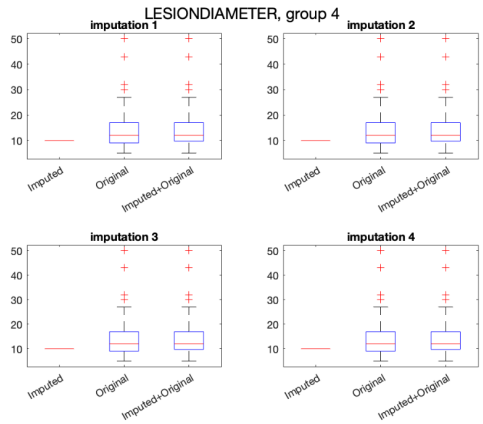
[26] V. N. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi, and V. Padma, "Study the influence of normalization/transformation process on the accuracy of supervised classification," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2020, pp. 729–735. DOI: 10.1109/ICSSIT48917.2020.9214160.

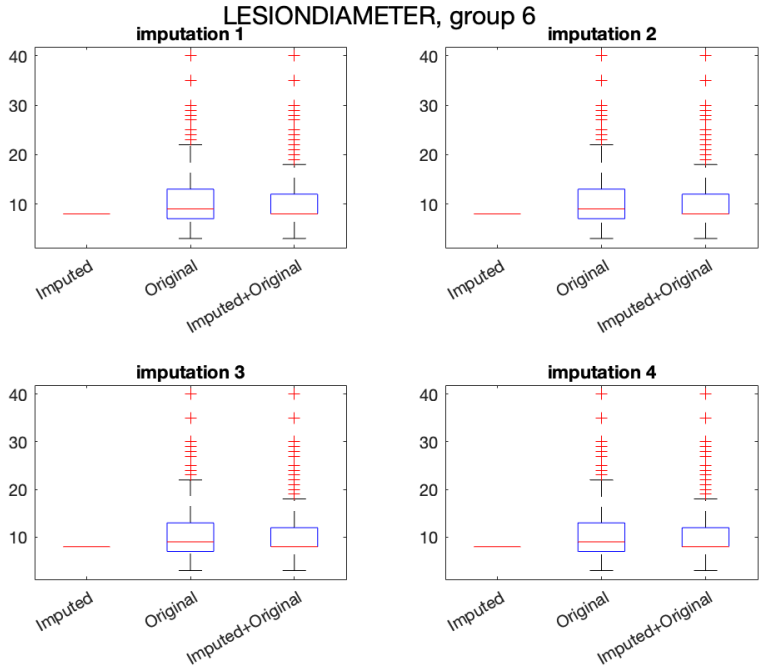
APPENDIX

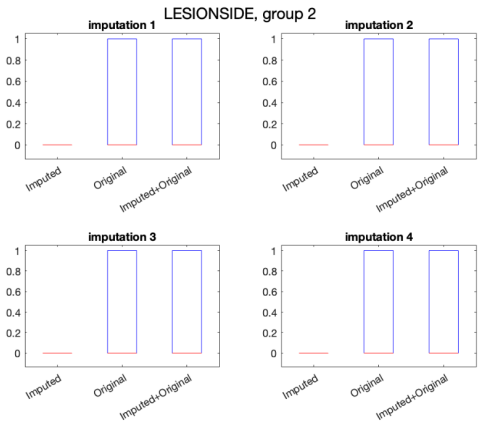
Boxplot, statitstical characterization of the imputation

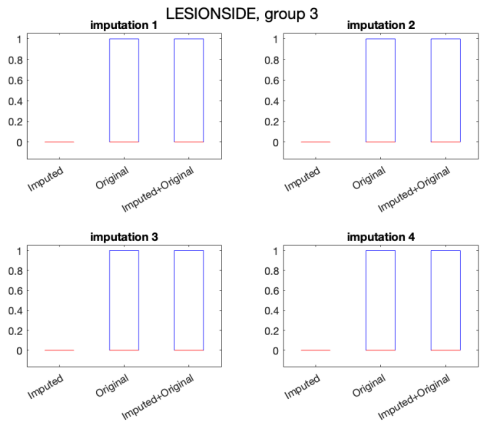


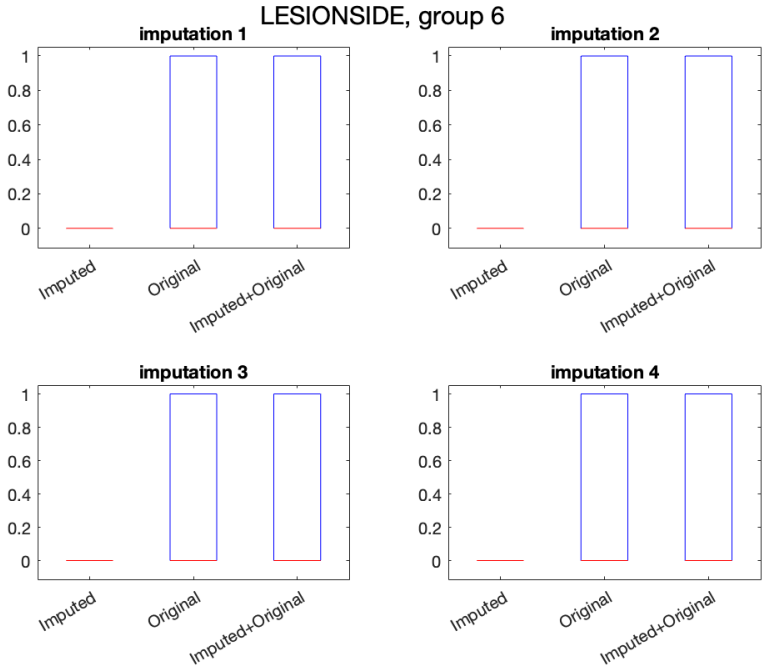


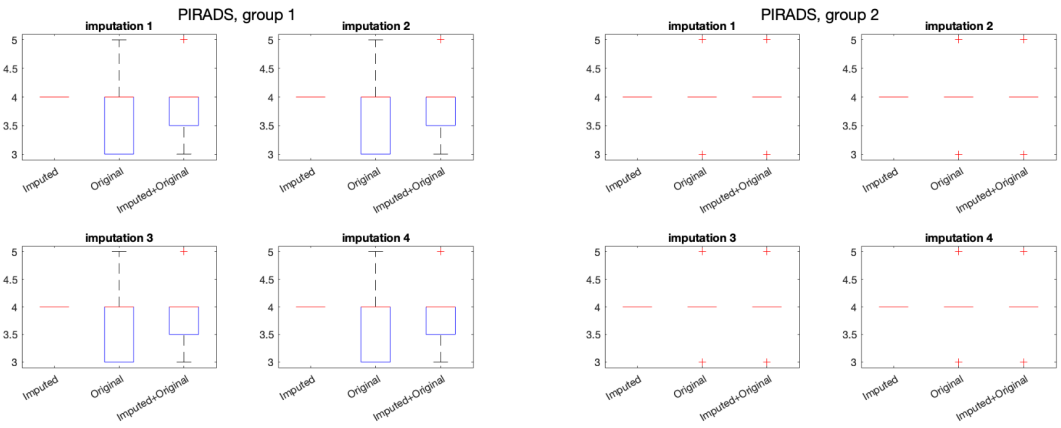


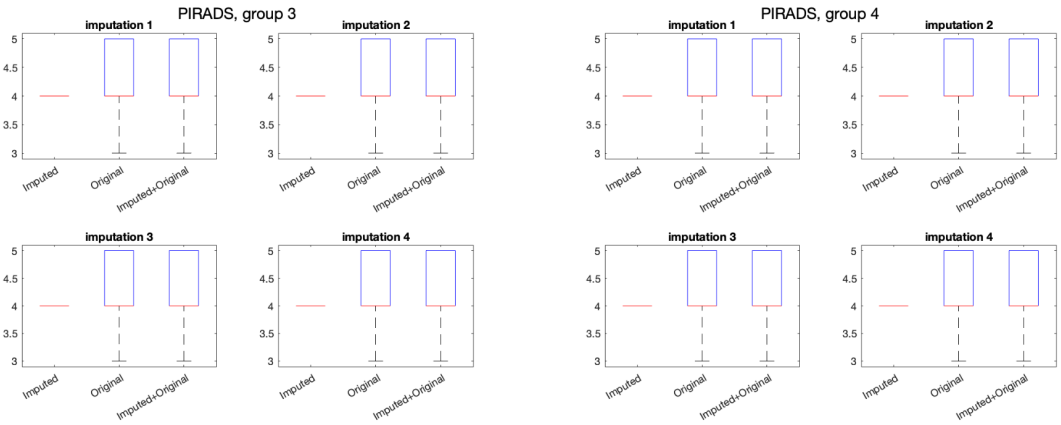

LESIONDIAMETER, group 6
imputation 1
imputation 2
imputation 3
imputation 4
40
30
20
10
Imputed
Original
Imputed+Original

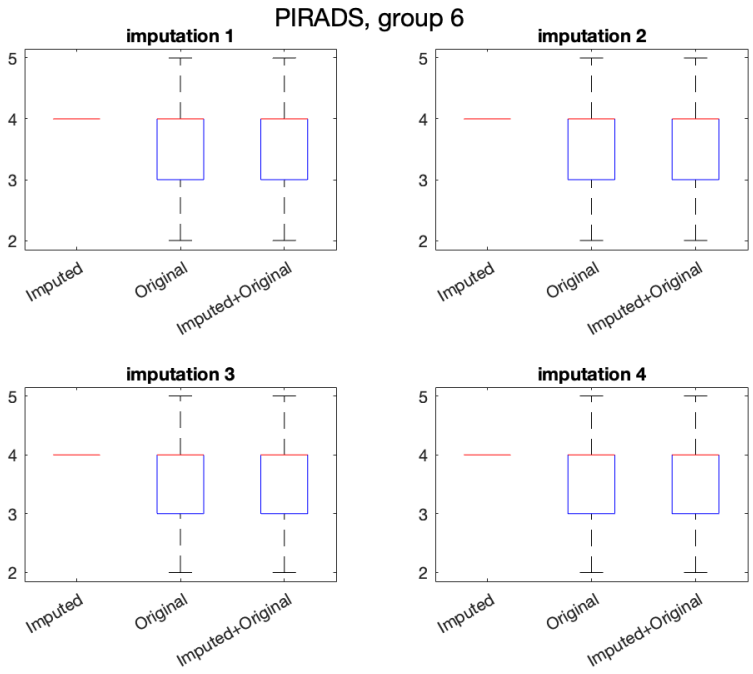

LESIONSIDE, group 2
imputation 1
imputation 2
imputation 3
imputation 4
1
0.8
0.6
0.4
0.2
0
Imputed
Original
Imputed+Original

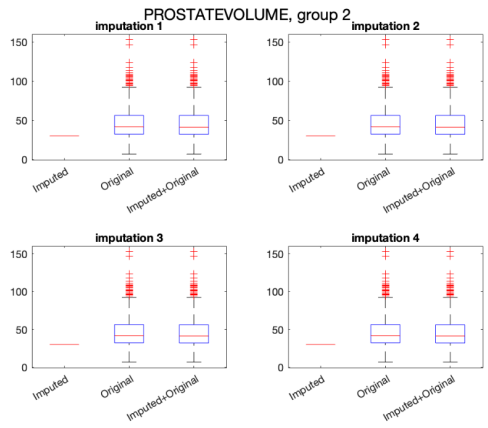

LESIONSIDE, group 3
imputation 1
imputation 2
imputation 3
imputation 4
1
0.8
0.6
0.4
0.2
0
Imputed
Original
Imputed+Original

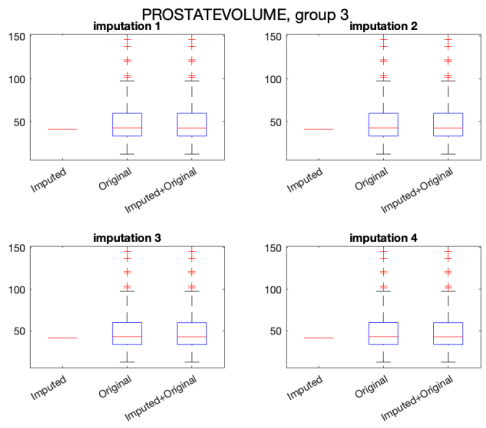

LESIONSIDE, group 6
imputation 1
imputation 2
imputation 3
imputation 4
1
0.8
0.6
0.4
0.2
0
Imputed
Original
Imputed+Original

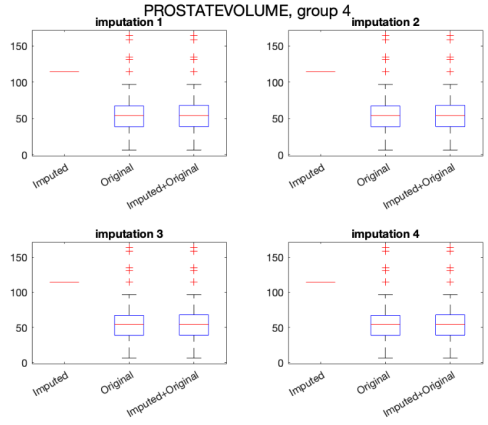

PIRADS, group 1
PIRADS, group 2
imputation 1
imputation 2
imputation 3
imputation 4
5
4.5
4
3.5
3
Imputed
Original
Imputed+Original

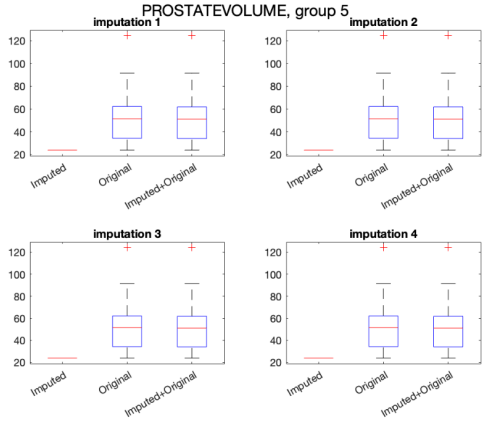

PIRADS, group 3
PIRADS, group 4
imputation 1
imputation 2
imputation 3
imputation 4
5
4.5
4
3.5
3
Imputed
Original
Imputed+Original

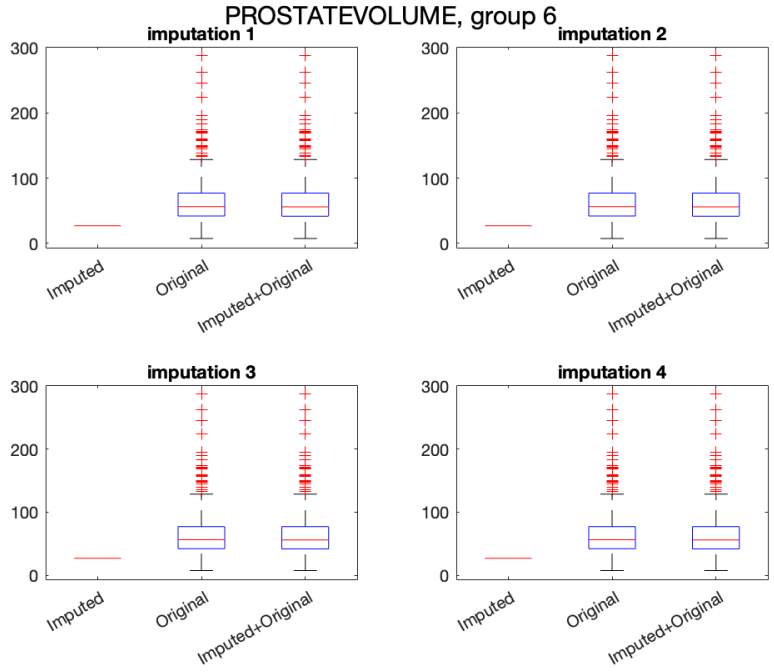

PIRADS, group 6
imputation 1
imputation 2
imputation 3
imputation 4
Imputed
Original
Imputed+Original

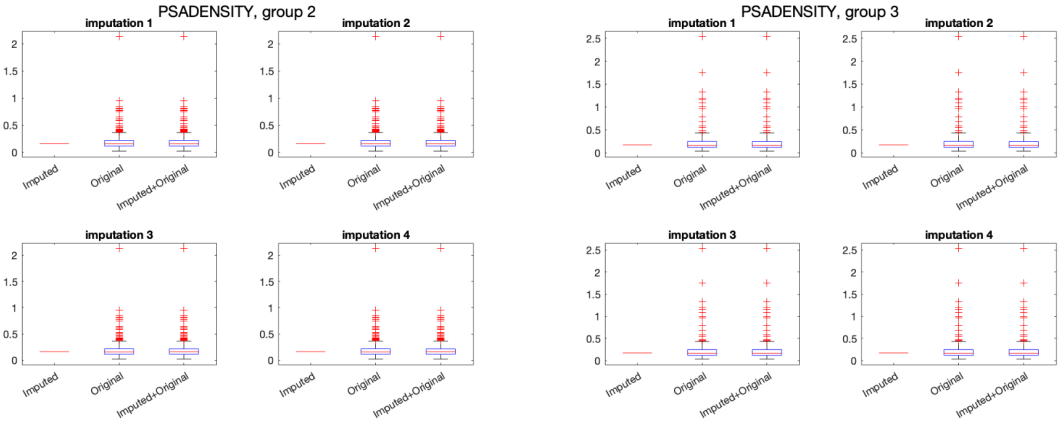

PROSTATEVOLUME, group 2
imputation 1
imputation 2
imputation 3
imputation 4
Imputed
Original
Imputed+Original

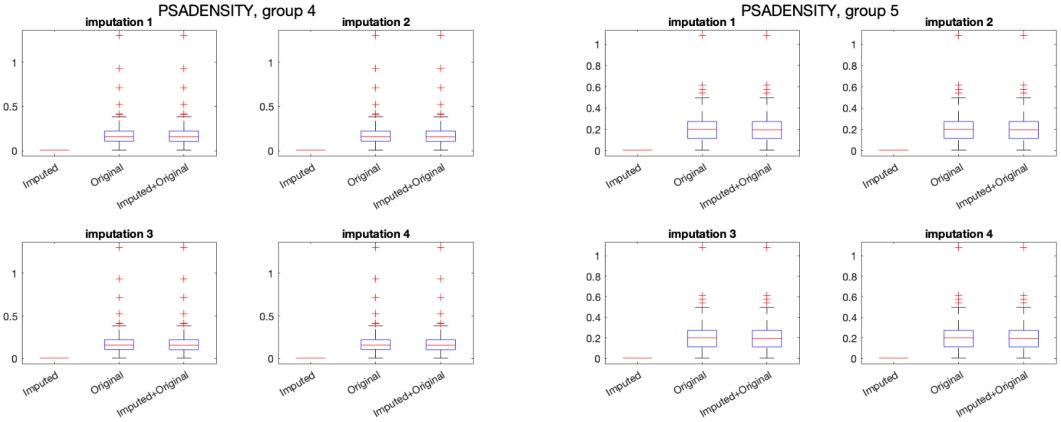

PROSTATEVOLUME, group 3
imputation 1
imputation 2
imputation 3
imputation 4
Imputed
Original
Imputed+Original

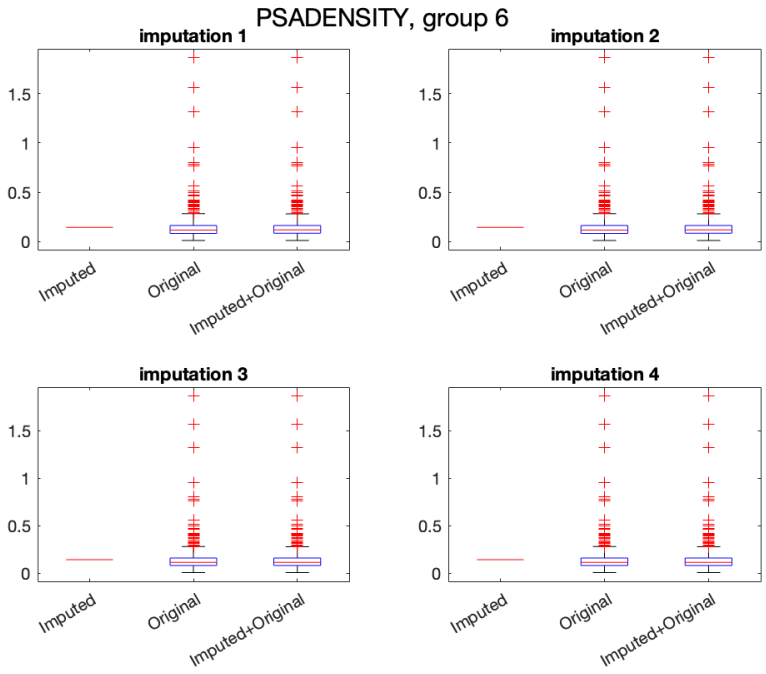

PROSTATEVOLUME, group 4
imputation 1
imputation 2
imputation 3
imputation 4
Imputed
Original
Imputed+Original

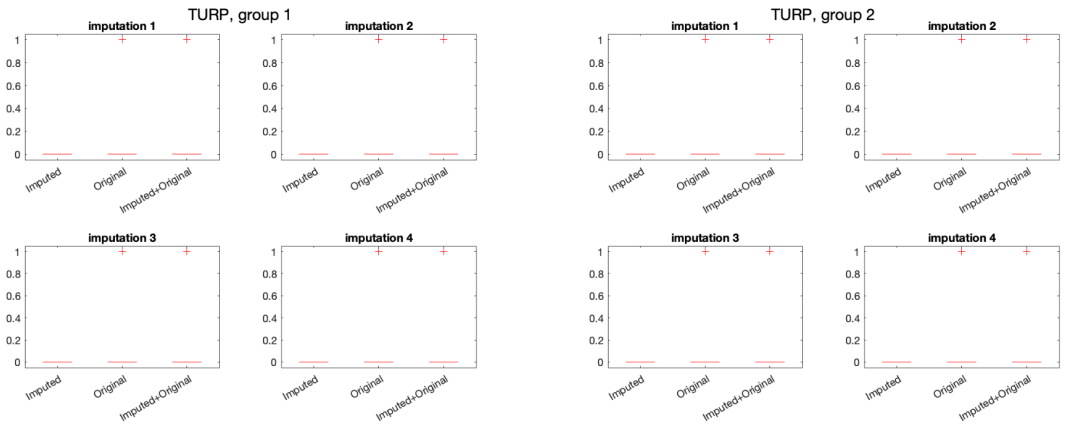

PROSTATEVOLUME, group 5
imputation 1
imputation 2
imputation 3
imputation 4
Imputed
Original
Imputed+Original

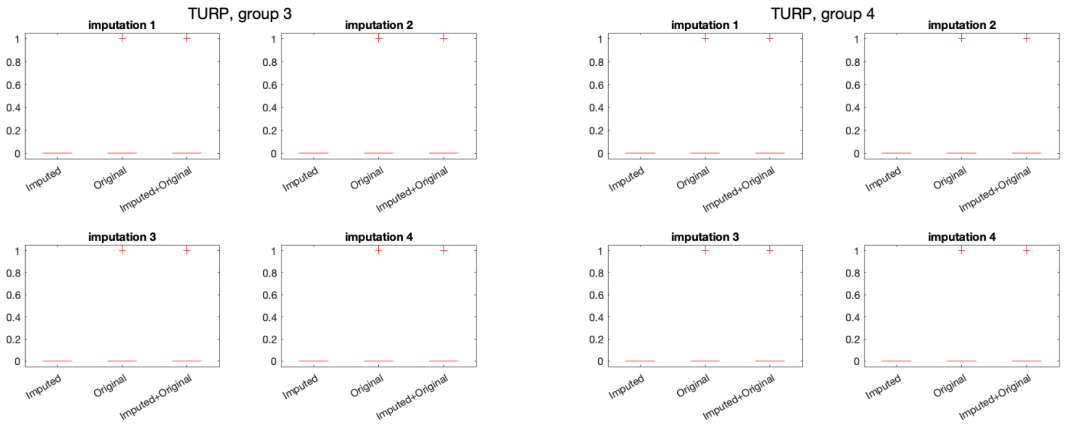

PROSTATEVOLUME, group 6
imputation 1
imputation 2
imputation 3
imputation 4
300
200
100
0
Imputed
Original
Imputed+Original

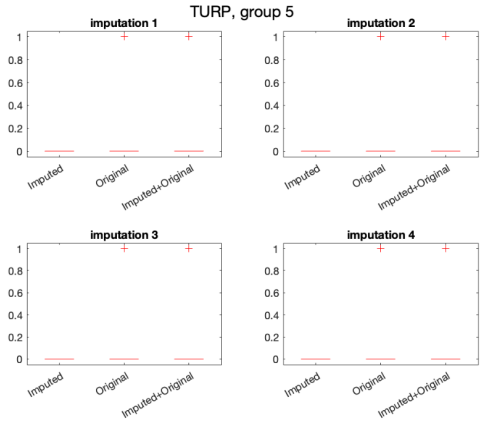

PSADENSITY, group 2
PSADENSITY, group 3
imputation 1
imputation 2
imputation 3
imputation 4
Imputed
Original
Imputed+Original

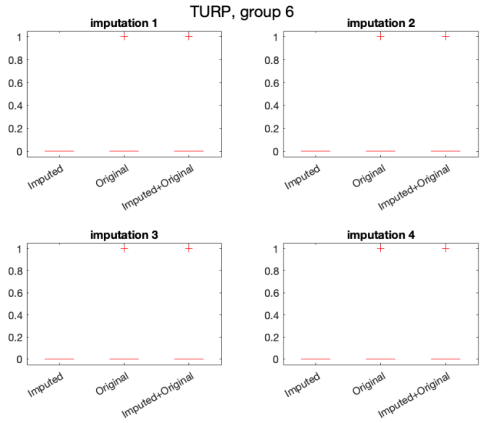

PSADENSITY, group 4
PSADENSITY, group 5
imputation 1
imputation 2
imputation 3
imputation 4
Imputed
Original
Imputed+Original


PSADENSITY, group 6
imputation 1
imputation 2
imputation 3
imputation 4
1.5
1
0.5
0
Imputed
Original
Imputed+Original


TURP, group 1
TURP, group 2
imputation 1
imputation 2
imputation 3
imputation 4
1
0.8
0.6
0.4
0.2
0
Imputed
Original
Imputed+Original

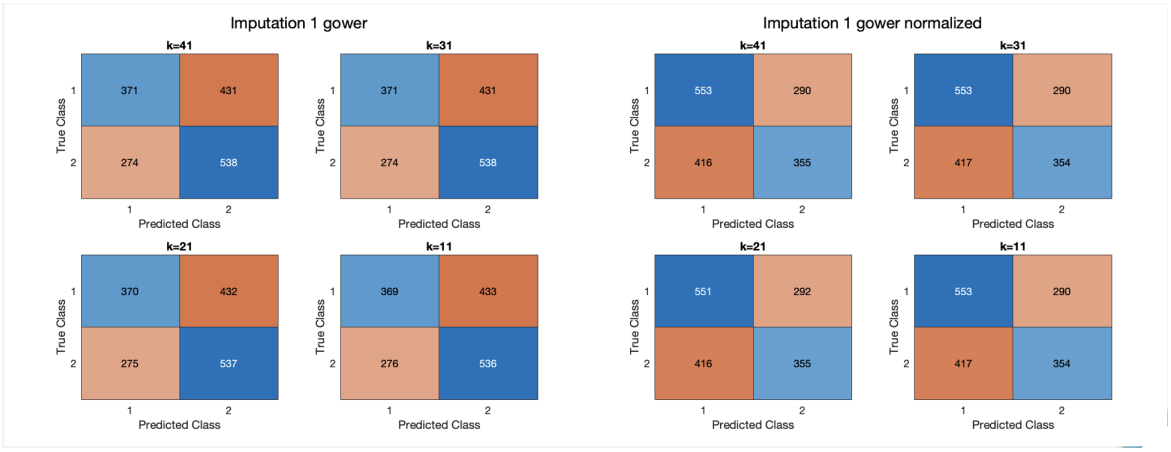

TURP, group 3
TURP, group 4
imputation 1
imputation 2
imputation 3
imputation 4
1
0.8
0.6
0.4
0.2
0
Imputed
Original
Imputed+Original

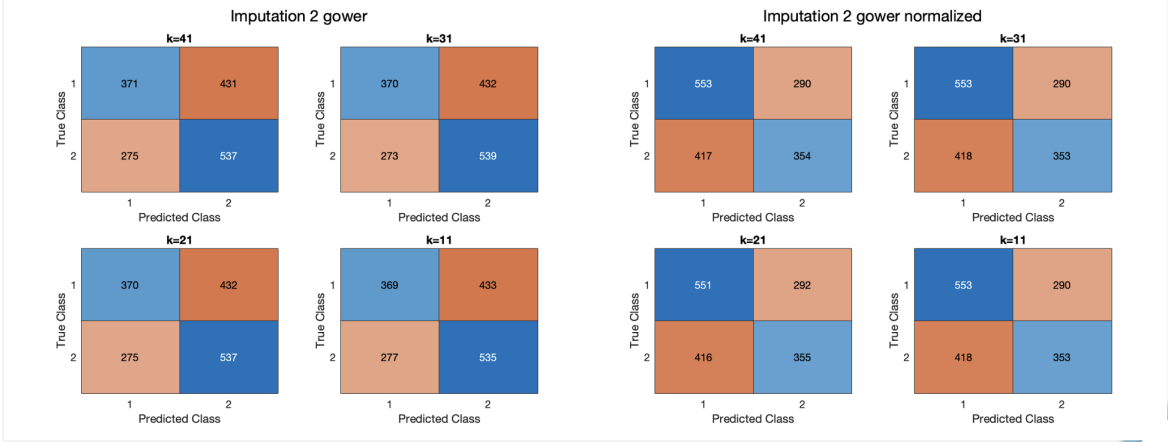


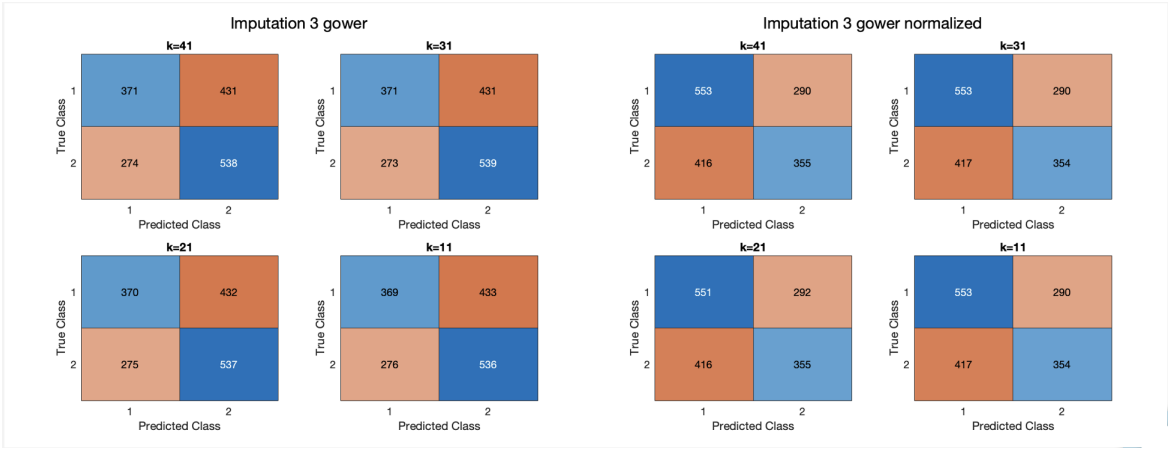


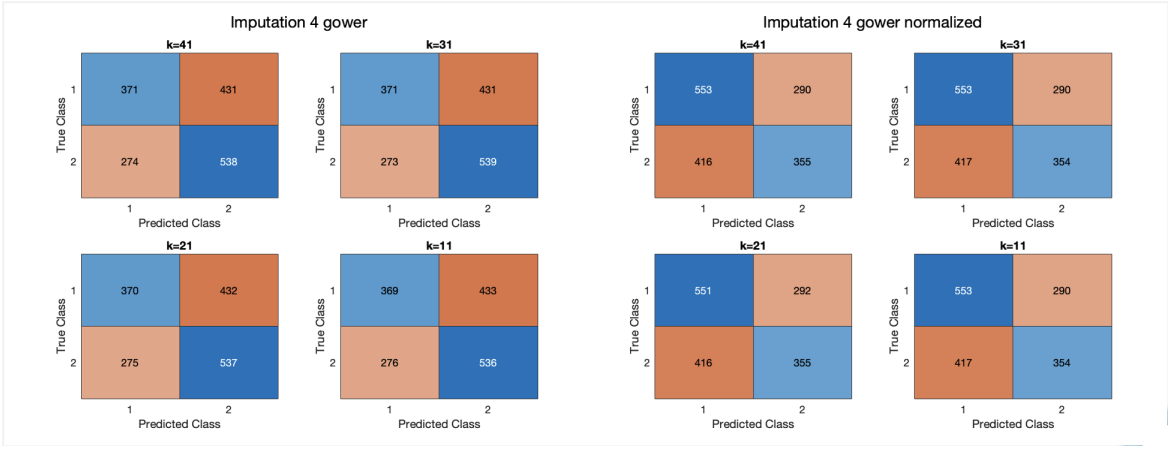
Confusion Matrix of the classifiers

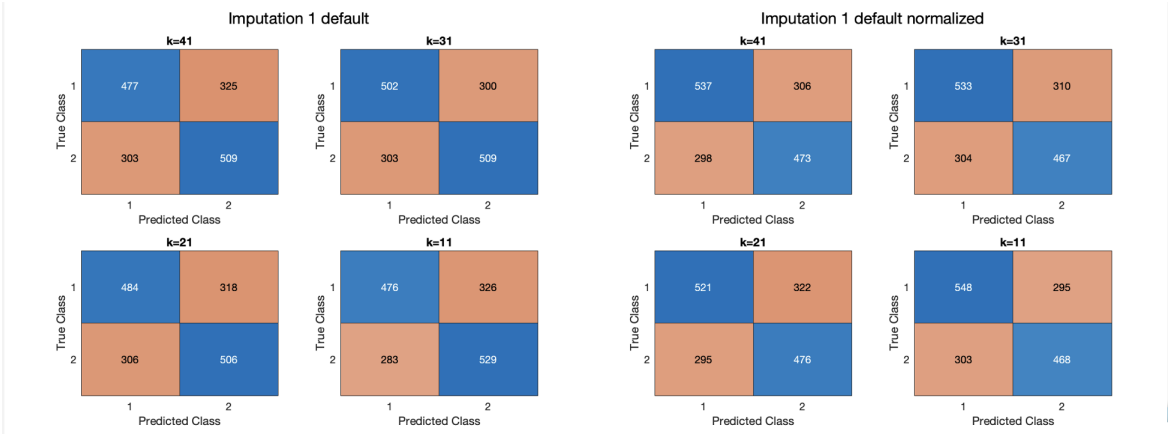
DATASET 1

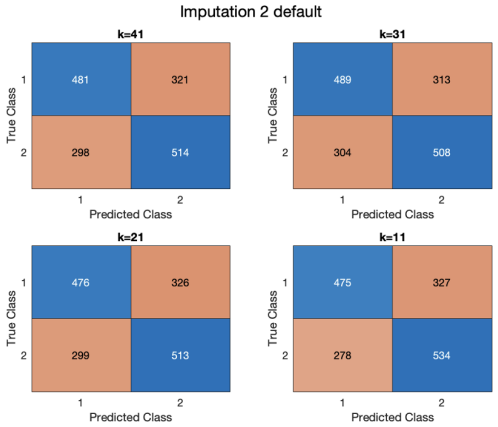


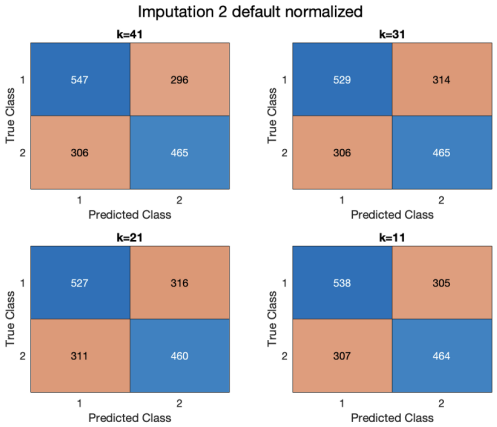


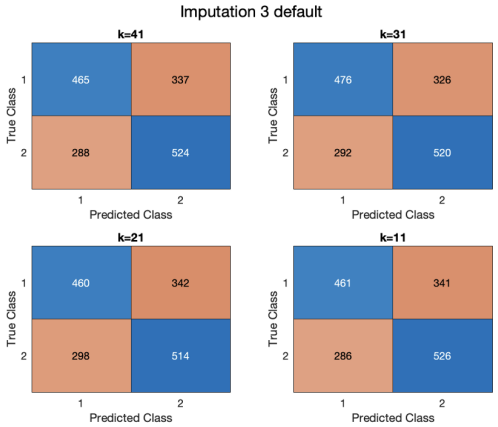

Imputation 3 gower
k=41
371
431
274
538
k=31
371
431
273
539
k=21
370
432
275
537
k=11
369
433
276
536
Imputation 3 gower normalized
k=41
553
290
416
355
k=31
553
290
417
354
k=21
551
292
416
355
k=11
553
290
417
354
True Class
Predicted Class

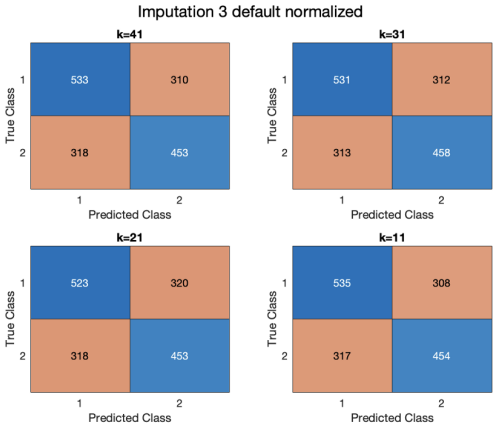

Imputation 4 gower
k=41
371
431
274
538
k=31
371
431
273
539
k=21
370
432
275
537
k=11
369
433
276
536
Imputation 4 gower normalized
k=41
553
290
416
355
k=31
553
290
417
354
k=21
551
292
416
355
k=11
553
290
417
354
True Class
Predicted Class

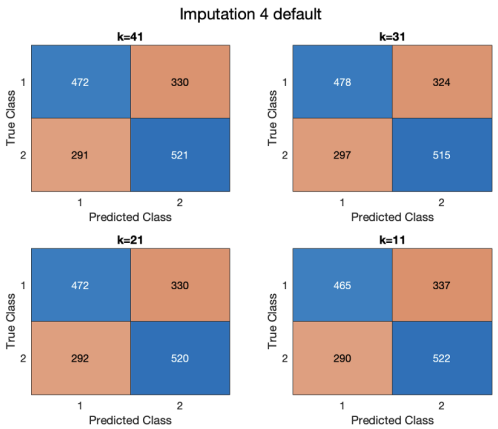

Imputation 1 default
k=41
477
325
303
509
k=31
502
300
303
509
k=21
484
318
306
506
k=11
476
326
283
529
Imputation 1 default normalized
k=41
537
306
298
473
k=31
533
310
304
467
k=21
521
322
295
476
k=11
548
295
303
468
True Class
Predicted Class

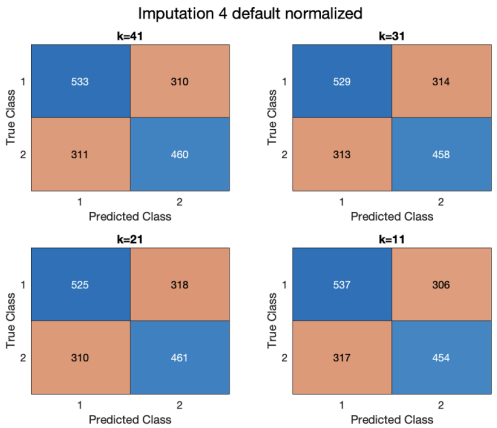

Imputation 2 default
k=41
481
321
298
514
k=31
489
313
304
508
k=21
476
326
299
513
k=11
475
327
278
534
True Class
Predicted Class


Imputation 2 default normalized
k=41
547
296
306
465
k=31
529
314
306
465
k=21
527
316
311
460
k=11
538
305
307
464
True Class
Predicted Class

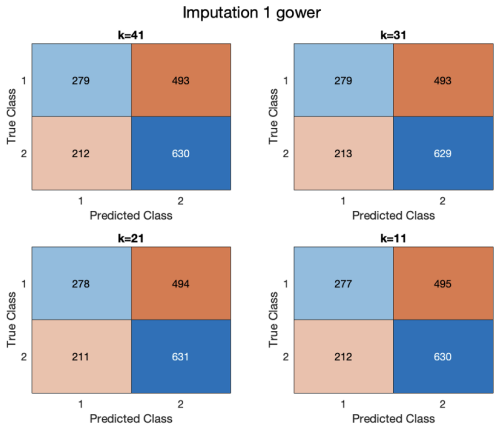

Imputation 3 default
k=41
465
337
288
524
k=31
476
326
292
520
k=21
460
342
298
514
k=11
461
341
286
526
True Class
Predicted Class

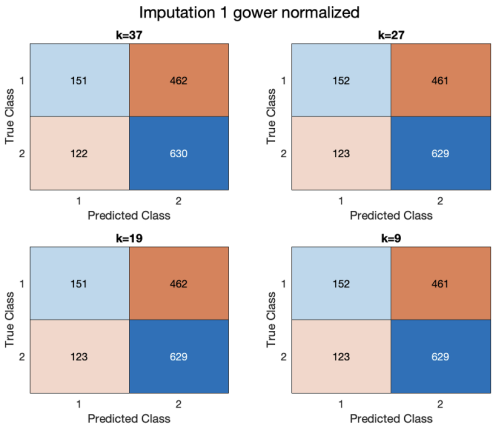

Imputation 3 default normalized
k=41
533
310
318
453
k=31
531
312
313
458
k=21
523
320
318
453
k=11
535
308
317
454
True Class
Predicted Class

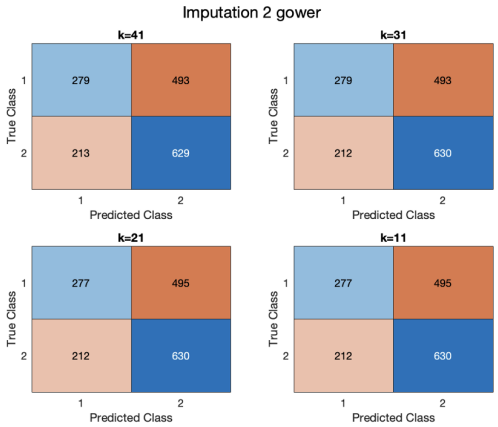

Imputation 4 default
k=41
472
330
291
521
k=31
478
324
297
515
k=21
472
330
292
520
k=11
465
337
290
522
True Class
Predicted Class

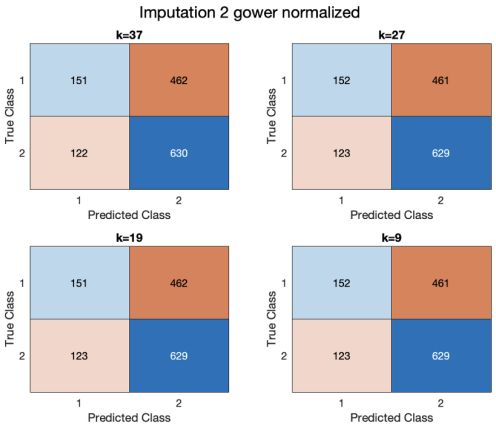

Imputation 4 default normalized
k=41
533
310
311
460
k=31
529
314
313
458
k=21
525
318
310
461
k=11
537
306
317
454
True Class
Predicted Class

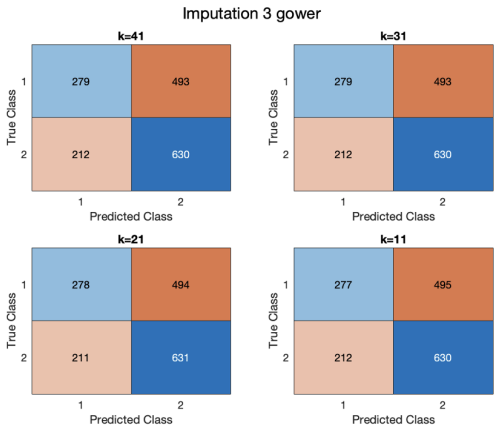
DATASET 2

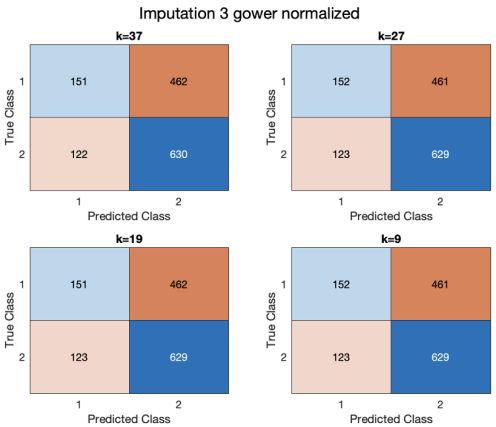

Imputation 1 gower
k=41
279
493
212
630
k=31
279
493
213
629
k=21
278
494
211
631
k=11
277
495
212
630
True Class
Predicted Class

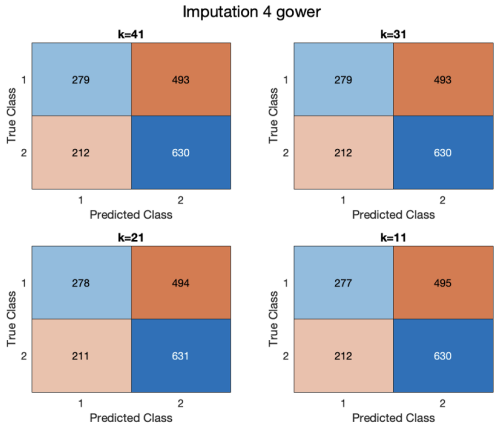

Imputation 1 gower normalized
k=37
151
462
122
630
k=27
152
461
123
629
k=19
151
462
123
629
k=9
152
461
123
629
True Class
Predicted Class

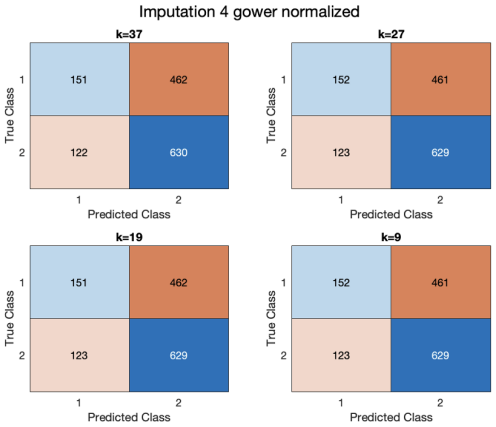

Imputation 2 gower
k=41
279
493
213
629
k=31
279
493
212
630
k=21
277
495
212
630
k=11
277
495
212
630
True Class
Predicted Class

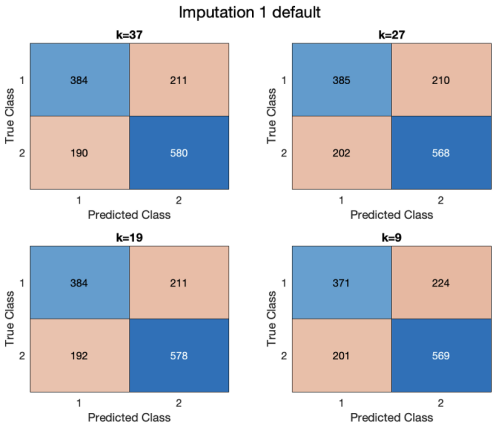

Imputation 2 gower normalized
k=37
151
462
122
630
k=27
152
461
123
629
k=19
151
462
123
629
k=9
152
461
123
629
True Class
Predicted Class

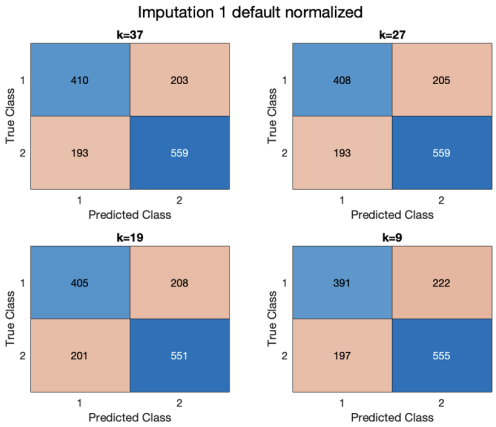

Imputation 3 gower
k=41
279
493
212
630
k=31
279
493
212
630
k=21
278
494
211
631
k=11
277
495
212
630
True Class
Predicted Class

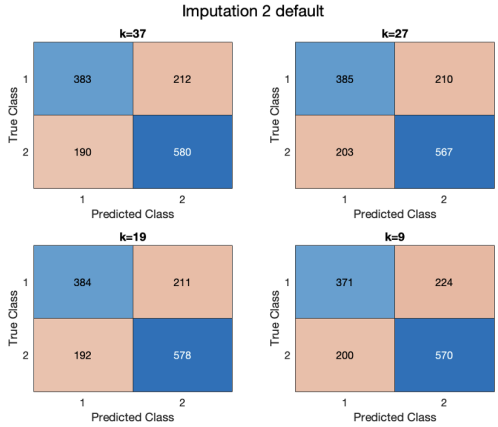

Imputation 3 gower normalized
k=37
151
462
122
630
k=27
152
461
123
629
k=19
151
462
123
629
k=9
152
461
123
629
True Class
Predicted Class

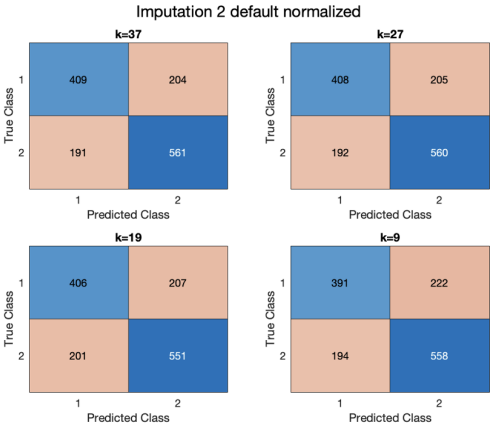

Imputation 4 gower
k=41
279
493
212
630
k=31
279
493
212
630
k=21
278
494
211
631
k=11
277
495
212
630
True Class
Predicted Class

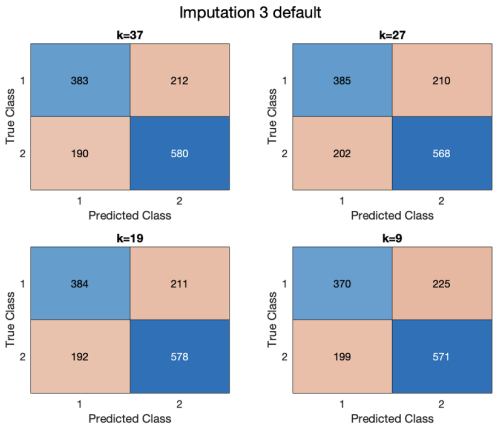

Imputation 4 gower normalized
k=37
151
462
122
630
k=27
152
461
123
629
k=19
151
462
123
629
k=9
152
461
123
629
True Class
Predicted Class

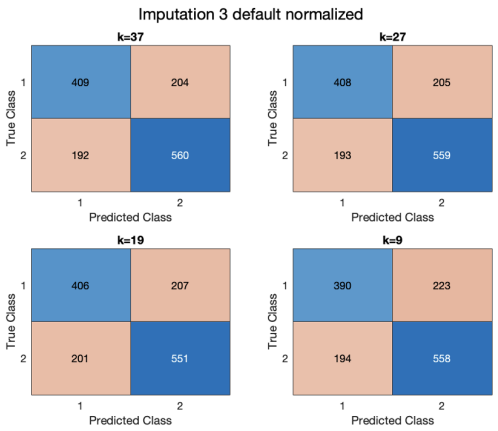

Imputation 1 default
k=37
384
211
190
580
k=27
385
210
202
568
k=19
384
211
192
578
k=9
371
224
201
569
True Class
Predicted Class

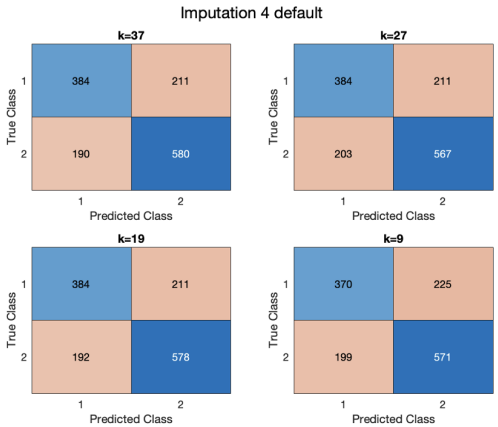

Imputation 1 default normalized
k=37
410
203
193
559
k=27
408
205
193
559
k=19
405
208
201
551
k=9
391
222
197
555
True Class
Predicted Class

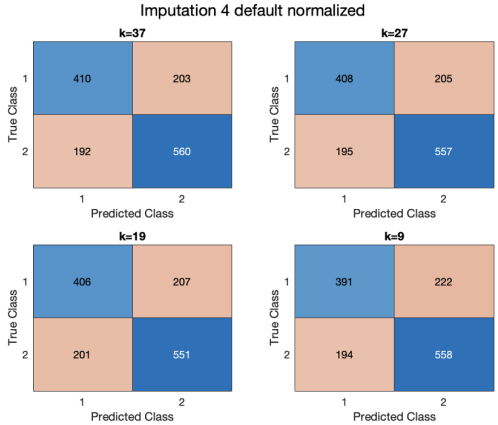

Imputation 2 default
k=37
383
212
190
580
k=27
385
210
203
567
k=19
384
211
192
578
k=9
371
224
200
570
True Class
Predicted Class


Imputation 2 default normalized
k=37
409
204
191
561
k=27
408
205
192
560
k=19
406
207
201
551
k=9
391
222
194
558
True Class
Predicted Class

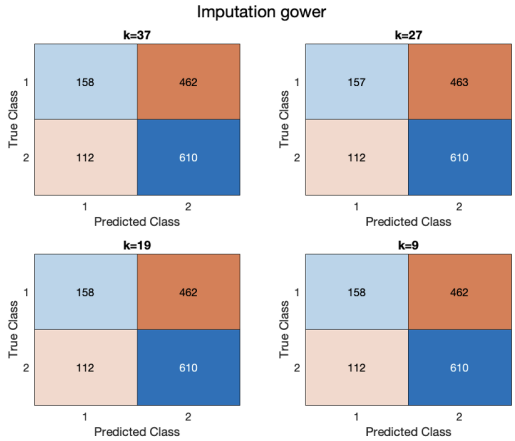

Imputation 3 default
k=37
383
212
190
580
k=27
385
210
202
568
k=19
384
211
192
578
k=9
370
225
199
571
True Class
Predicted Class

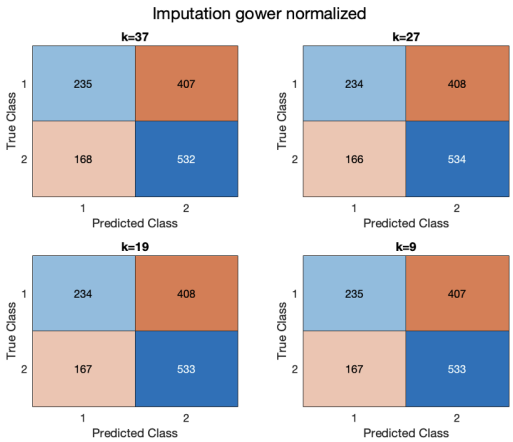

Imputation 3 default normalized
k=37
409
204
192
560
k=27
408
205
193
559
k=19
406
207
201
551
k=9
390
223
194
558
True Class
Predicted Class

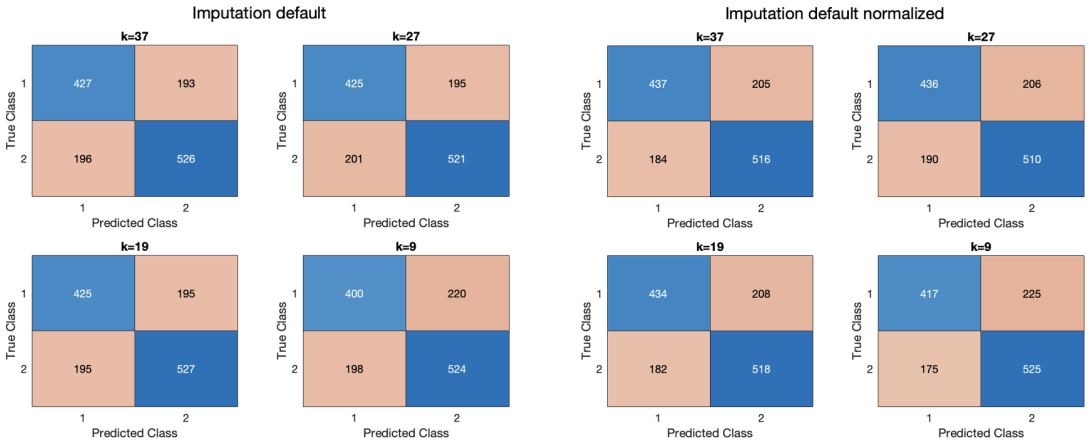

Imputation 4 default
k=37
384
211
190
580
k=27
384
211
203
567
k=19
384
211
192
578
k=9
370
225
199
571
True Class
Predicted Class


Imputation 4 default normalized
k=37
410
203
192
560
k=27
408
205
195
557
k=19
406
207
201
551
k=9
391
222
194
558
True Class
Predicted Class

DATASET 3


Imputation gower
k=37
158
462
112
610
k=27
157
463
112
610
k=19
158
462
112
610
k=9
158
462
112
610
True Class
Predicted Class


Imputation gower normalized
k=37
235
407
168
532
k=27
234
408
166
534
k=19
234
408
167
533
k=9
235
407
167
533
True Class
Predicted Class


Imputation default
k=37
427
193
196
526
k=27
425
195
201
521
k=19
425
195
195
527
k=9
400
220
198
524
Imputation default normalized
k=37
437
205
184
516
k=27
436
206
190
510
k=19
434
208
182
518
k=9
417
225
175
525
True Class
Predicted Class
1
2