

Degree in Computer Science
Computation

Final Degree Thesis

Analysis of the influence of sex in diagnostic classification of Parkinson's disease based on non-motor manifestations by means of machine learning methods

Author

Ander Barrio Campos

2022

Degree in Computer Science
Computation

Final Degree Thesis

Analysis of the influence of sex in diagnostic classification of Parkinson's disease based on non-motor manifestations by means of machine learning methods

Author

Ander Barrio Campos

Director

Olatz Arbelaitz Gallego

Abstract

Parkinson's disease (PD) is the second most common neurodegenerative disorder, after Alzheimer's disease. In the early stages of the disease, when motor symptoms have not yet manifested themselves, the accuracy of making a correct diagnosis is currently very limited. This work aims to analyse the influence of sex in diagnostic classification of Parkinson's disease based on non-motor symptoms by using machine learning methods. These symptoms have been evaluated in 490 subjects with PD and 197 healthy control subjects.

Contents

Abstract	i
Contents	iii
List of Figures	vii
List of Tables	ix
1 Introduction	1
2 Document of the Project Objectives	3
2.1 Planning	3
2.1.1 Risk and countermeasures	3
2.2 Monitoring	5
2.2.1 Deviation from the plan	6
3 Background	7
3.1 Parkinson Disease	7
3.2 Machine Learning and sex analysis in health	8
3.3 Parkinson diagnosis based on non-motor symptoms	10

4	Theoretical Concepts	13
4.1	Machine Learning methods	13
4.1.1	SVM	13
4.1.2	Multilayer perceptron	16
4.1.3	XGBoost	22
4.2	Fairness metrics	26
4.2.1	Demographic parity	26
4.2.2	Bias score	27
4.3	Evaluation	27
4.4	SHAP	28
5	Technology	31
6	Project development	35
6.1	Methodology	35
6.2	Databases	36
6.2.1	Biocruces	36
6.2.2	PPMI	36
6.2.3	FPT	38
6.2.4	FPI	39
6.2.5	FPTSex	40
6.2.6	FPISex	40
6.3	Results	41
6.3.1	PD/HC classification	41
6.3.2	Sex classification	48

7	Conclusions	59
7.1	Personal conclusions	60
7.2	Future work	60
 Appendices		
A	Other metrics figures	63
 Bibliografia		
		71

List of Figures

2.1	Work Breakdown Structure.	4
4.1	SVM hyperplane.	15
4.2	SVM Kernel.	16
4.3	Perceptron.	17
4.4	Most widely used activation functions.	19
4.5	Multilayer perceptron.	20
4.6	Decision Tree classification.	23
4.7	Gradient Boosting.	24
4.8	SHAP values impact on model output.	29
6.1	SDMT.	37
6.2	MoCA.	38
6.3	PD/HC class distribution.	42
6.4	Accuracies for FPT.	43
6.5	Accuracies for FPI.	44
6.6	Confusion matrices for FPT.	45
6.7	Confusion matrices FPI.	46
6.8	Demographic parity for PD/HC classification.	47
6.9	Bias towards PD.	48

6.10	XGBoost ranking for PD/HC classification.	49
6.11	SHAP in FPT.	50
6.12	SHAP in FPI.	51
6.13	Sex class distribution.	52
6.14	Demographic parity for PD/HC classification.	53
6.15	Demographic parity for PD/HC classification.	54
6.16	Confusion matrices for PD.	54
6.17	Confusion matrices for HC.	55
6.18	Demographic parity for sex classification.	56
6.19	Bias towards being a man.	56
6.20	XGBoost ranking for sex classification.	57
A.1	F-measure for FPT and FPI.	64
A.2	Recall for FPT and FPI.	65
A.3	Precision for FPT and FPI.	66
A.4	F-measure for FPTSex and FPISex.	67
A.5	Recall for FPTSex and FPISex.	68
A.6	Precision for FPTSex and FPISex.	69

List of Tables

2.1	Tasks and planned time spans.	5
2.2	Tasks and real time spans.	6
6.1	Distribution of BIO	36
6.2	Distribution of FPT	38
6.3	Distribution of FPTWomen	39
6.4	Distribution of FPTMen	39
6.5	Distribution of FPTSex	40
6.6	Distribution of FPTSexPD	40
6.7	Distribution of FPTSexHC	40

1. CHAPTER

Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disorder, after Alzheimer's disease [2]. PD has a prevalence of approximately 0.5 to 1 percent among persons 65 to 69 years, rising to 1 to 3 percent among persons 80 years of age and older. The diagnosis is made clinically, although other disorders with prominent symptoms and signs of parkinsonism (resting tremor, bradykinesia, rigidity and postural instability)The diagnosis is made clinically, although other disorders with prominent symptoms and signs of parkinsonism. The pathological diagnosis is noticed by the loss of neurons in the substantia nigra [47]. Nigro-striatal dopaminergic neurones are particularly well suited to the study of neurotransmitter release from dendrites in the central nervous system. Their cell bodies are mainly concentrated in the pars compacta and their long, ramified and varicose dendrites extend throughout the pars reticulata of the substantia nigra [14].

There are two main types of symptoms related with PD. Motor symptoms and non-motor symptoms. On the one hand, the motor symptoms begin to manifest themselves slowly, without being very aggressive at the beginning of the disease, on one side of the body. Eventually they begin to affect both sides of the body. When someone asks or talks about Parkinson's, the first thought that comes to mind are the tremors. These tremors affect the hands, arms, legs, jaw and face. However, there are more bodily alterations due to this disease. These alterations are related to stiffness in the arms, legs and trunk, slowness of movement and problems with balance and coordination. On the other hand, non-motor symptoms in PD involve a multitude of functions including sleep-wake cycle regulation, cognitive function, regulation of mood and hedonistic tone, autonomic nervous system function as well as sensory function and pain perception [50]. In this work, we will

take into account the non-motor symptoms.

Parkinson's disease seems to occur more commonly in men than women [56] based primarily on studies of death rates and prevalence. In recent years, several population based incidence studies of Parkinson's disease that included sex data have been conducted in a variety of populations around the world. A significantly higher incidence rate of Parkinson's disease was found among men with the relative risk being 1.5 times greater in men than women. Possible reasons for this increased risk of Parkinson's disease in men are toxicant exposure, head trauma, neuroprotection by oestrogen, mitochondrial dysfunction, or X linkage of genetic risk factors [70].

Males and females have different patterns of illness and different life spans [34]. Understanding the bases of these sex-based differences is important to developing new approaches to prevention, diagnosis, and treatment [49].

In recent years, supervised learning and the healthcare world [57] have gone hand in hand. This fact is something very beneficial for society since it allows us to analyze medical data in a way never seen before.

Artificial Intelligence (AI) can potentially impact many aspects of human health, from basic research discovery to individual health assessment. It is critical that these advances in technology broadly benefit diverse populations from around the world. This can be challenging because AI algorithms are often developed on non-representative samples and evaluated based on narrow metrics [77].

Numerous researchers have identified ways in which non-health related machine learning can exacerbate existing social inequalities by reflecting and amplifying existing race, sex and other biases. Health care is not immune to pernicious bias. The health data on which algorithms are trained are likely to be influenced by many facets of social inequality, including bias toward those who contribute the most data. For example, algorithms for predicting whether an individual should receive a surgery may be biased toward those who are able to access and afford the procedure [68].

2. CHAPTER

Document of the Project Objectives

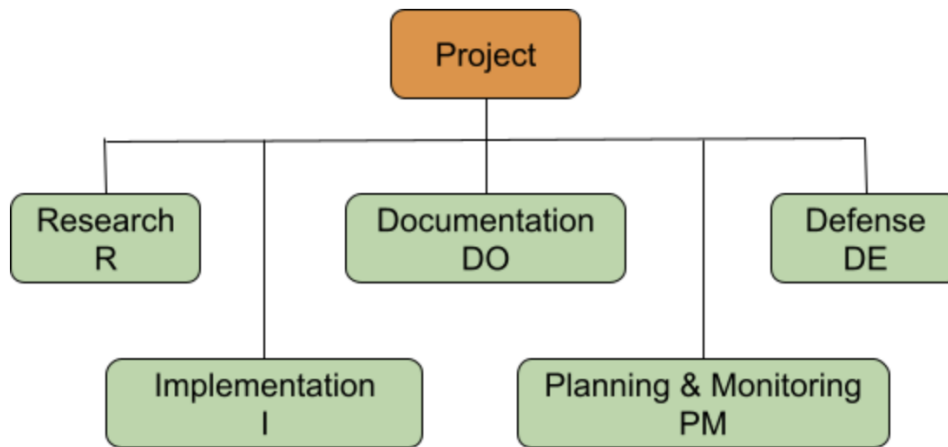
Parkinson's disease (PD) is the second most common neurodegenerative disorder, after Alzheimer's disease. In the early stages of the disease, when motor symptoms have not yet manifested themselves, the accuracy of making a correct diagnosis is currently very limited. This work aims to analyse the influence of sex in diagnostic classification of Parkinson's disease based on non-motor symptoms by using machine learning methods. These symptoms have been evaluated in 490 subjects with PD and 197 healthy control subjects.

2.1 Planning

The concept of planning a project is related to complete a project that has been divided into defined stages in a certain time frame. These stages have been making a research about the topic, implementation of different machine learning methods, documentation of the project, planning and monitoring the proposed tasks and preparing the defense. Project planning is important at every phase of a project. The work breakdown structure and the planned spans for each task can be seen in figure 2.1 and table 2.1.

2.1.1 Risk and countermeasures

R1. Difficulties in meeting with the tutor.



2.1 Figure: Work Breakdown Structure.

It is very difficult for both the student and the tutor to maintain daily communication. That is why if urgent questions need to be answered, they should wait until the next scheduled meeting.

To manage this risk it is necessary to clarify as many doubts and ambiguities as possible at each scheduled meeting.

R2. Combining the subjects and the project.

The fact that the project is carried out during the second term of the final year means that the hours devoted to passing the subjects and those devoted to carrying out the project tasks have to be combined. This means less time for both the subjects and the project and can lead to a backlog of assignments.

In order to avoid this risk it is necessary to manage the available time properly and to clearly prioritise between tasks.

R3. Dissatisfaction with the work carried out.

It is possible that the tutor does not agree with the development of the project and it may not be possible to hand it in.

That is why it is advisable to hold weekly meetings to control the progress made. In this way, the mistakes made can be corrected for the next meeting and so on and we will avoid last minute surprises.

Code	Task	Estimation
PM1	Planning the project	5h
PM2	Monitoring the development	5h
R1	Finding additional bibliography	10h
R2	Understanding the most important papers	30h
R3	Selection of the classification models	5h
R4	Explore the given databases	10h
R5	Finding extra bibliography for the report	25h
I1	Creating the different levels of databases	10h
I2	Implementation of fairness metrics	5h
I3	Implementation of classification models	25h
I4	Experimentation	10h
I5	Testing and correction	5h
DO1	Extraction of the conclusions from I4	20h
DO2	Graphics creation for the report	15h
DO3	Writing of the report	80h
DE1	Making the presentation for the defense	20h
DE2	Preparing the oral defense	20h
-	Total	300h

2.1 Table: Tasks and planned time spans.

R4. Changes in the middle of the project.

Although unlikely, there is always a small chance that the objective of the project may change. Therefore, all the tasks and dedications set at the beginning will have to be changed, making the loss of time very high.

This risk cannot be afforded in any way as the changes can be sudden.

2.2 Monitoring

In general, the development of the project has been smooth. The weekly meetings between the tutor and the student have been key to achieving the objectives set. In these meetings, emphasis was placed on reviewing the work done during the week and ideas and strategies were put forward for the following week's work.

Being able to work on the project during the first months has been combined with trying to pass the remaining subjects to finish the degree. That is why the work accumulated in

the first months was remarkable and the effort, dedication and attention paid to the project were not remarkable. However, after having passed the corresponding subjects, the total focus has been on the realisation of the project.

2.2.1 Deviation from the plan

The following table shows the difference in time between how much each defined task actually cost to perform and how much was originally estimated. The tasks related to defense (making the presentation for the defense of the project and preparing the oral defense) have not been recorded yet. This is because the documentation of the project must be uploaded earlier than the slides for the oral presentation. Thus, an estimation time have been added for these two tasks.

Code	Task	Estimation	Deviation
PM1	5h	5h	+0h
PM2	5h	5h	+0h
R1	10h	10h	+0h
R2	30h	32h	+2h
R3	5h	2h	-3h
R4	10h	10h	+0h
R5	25h	27h	+2h
I1	10h	10h	+0h
I2	5h	2h	-3h
I3	25h	25h	+0h
I4	10h	10h	+0h
I5	5h	7h	+2h
DO1	20h	20h	+0h
DO2	15h	15h	+0h
DO3	80h	90h	+10h
DE1	20h	20h	+0h
DE2	20h	20h	+0h
Total	300h	300h	+9h

2.2 Table: Tasks and real time spans.

3. CHAPTER

Background

3.1 Parkinson Disease

Parkinson's disease (PD) is the second most common neurodegenerative disorder, after Alzheimer's disease [2]. It is characterized clinically by parkinsonism (resting tremor, bradykinesia, rigidity and postural instability) and pathologically by the loss of neurons in the substantia nigra [47]. The disease was first described by Dr. James Parkinson in 1817 as a "shaking palsy"[22]. It is a chronic, progressive neurodegenerative disease characterized by both motor and non-motor features. The disease has a significant clinical impact on patients, families, and caregivers through its progressive degenerative effects on mobility and muscle control.

Looking at developed countries there will be over 1 million PD patients in US alone by 2020, a number larger than total patients with multiple sclerosis, muscular dystrophy and Amyotrophic Lateral Sclerosis (ALS) [66]. Worldwide nearly 10 million people are estimated to be affected by PD. PD prevalence is likely to be twofold by 2030 [20]. In these estimates, one should also factor in the lesser amount of awareness and diagnosis in developing countries.

There exists motor and non-motor symptoms. This work has focused solely and exclusively on non-motor symptoms of PD. Therefore, motor symptoms are of no importance this time. Regarding to non-motor symptoms in PD [32], many of which antedating motor dysfunction and representing a preclinical phase spanning 20 or more years, are linked to widespread distribution of α -synuclein pathology not restricted to the dopaminergic nigros-

striatal system that is responsible for core motor features of PD. The pathologic substrate of non-motor manifestations such as olfactory, autonomic (gastrointestinal, urogenital, cardia, respiratory), sensory, skin, sleep, visual, neuropsychiatric dysfunctions (cognitive, mood, dementia), and others are critically reviewed.

No cure for Parkinson's has been found, although there are medicines that help to improve the symptoms. These symptoms can be managed with several different drugs. The drugs used to treat PD either boost the levels of dopamine in the brain or mimic the effects of dopamine [59]. Many of people with Parkinson's disease will therefore be substantially dependent on clinical intervention. The requisite physical visits to the clinic for monitoring and treatment are difficult for many people with Parkinson's disease [15].

There is currently no diagnostic test for Parkinson's disease. Professionals use the patient's medical history and a neurological examination to make the diagnosis. The average age at which a person develops the disease is 60 years. There are cases where it manifests earlier, although this is rare.

3.2 Machine Learning and sex analysis in health

The incorporation of machine learning into the world of healthcare [10] instills hope that the global health system can be improved. However, there are some ethical challenges that raise great concern. One clear example is that algorithms have the possibility to mimic human biases in decision making.

The following section is a summary of the main articles that have inspired me to carry out this project applying the techniques and strategies that have been used.

- In [67] Wang et al.(2021) investigated survival prediction using machine learning algorithms among patients that suffer from lung cancer. The database that has been used in the paper consists mostly of white population and the observations of this study may serve as a good reflection of the sex-disparity in white lung cancer patients. The methods used for classification were Naive Bayes, Decision Tree, Random Forest, XGBoost, KNN, Logistic Regression and Support Vector Machine. When evaluating the classifiers, metrics as Accuracy, F1 score and sensitivity specificity were applied.

Two statistical tests, Chi-squared test (Chi2test) and t-test (Ttest), were used to evaluate the association between the category or continuous features and sex infor-

mation. The Cox regression was used to perform univariate survival analysis of the individual features. All the prognostic factors derived from the univariate analysis were collected for the multivariate analysis. However, all these techniques will not be applicable to our case.

The authors introduce separated models for the complete sample, female sample and male sample. In the end, they obtain feature importance for each of the samples in XGBoost and discuss about the obtained results.

- Garnica-Caparrós et al. (2021) presented a methodology that uses multiple supervised classification algorithms to predict player sex from games actions [29]. The study was able to determine important factors that differentiate each sex performance. Each model was evaluated to ensure a fitted classification and the author made use of explainability methods to induce significant differences identified by the models. In detail, three representative binary classification models were trained. These models were a logic-based Decision Tree algorithm, a probabilistic Logistic Regression model and a multi-level Neural Network perceptron.

The trained Neural Network model was used to compute SHAP (Shapley Additive Explanations) based on Shapley Values. The basic idea of these values was to map the model's prediction as a payout and the features as the game players. Shapley values then told the authors how to distribute the payout among the features equitably. Computing Shapley Values required a lot of computing power and time. That is why, in most real-world cases, only the estimated Shapley Values are feasible.

SHAP estimates the contribution of each feature value to the prediction. The results obtained showed a high prediction accuracy and a shared conclusion for the three used methods.

- Niklason et al. (2021) [45] showed how Cannabis Use Disorder (CUD) had been linked to environmental, personality, mental health, neurocognitive and neurobiological risk factors. Many studies had revealed sex importance differences in CUD. However, the relative importance of these complex factors by sex had not been described.

A data-driven examination of sex differences in CUD in a community sample of young adults was conducted. The sample had more than a thousand individuals and more than a half were female.

Regarding to machine learning methods, the one selected was XGBoost. SHAP (SHapley's Additive exPlanations) was used as a novel factor ranking tool. SHAP

provided an explanation model that computed the unique and additive importance of each model feature (predictive factor) in determining the final classification result. The impact of each feature on the output of the model is defined as the change in model output when the feature is known, as opposed to unknown. The performance of each model was quantified by using the Area Under the Curve of the Receiver (AUC).

To determine which factors drove the performance of the best performing classification models, SHAP was used to estimate the relative importance of all factors. To determine which factors consistently classified increased cannabis use levels and dependence, the median rank of each factor across all models was computed. Interaction effects to identify sex-specific factors that contribute to classifying cannabis dependence were examined as well. There was a special attention on the models predicting cannabis dependence.

The results were able to successfully classify both cannabis dependence and cannabis use levels. Previously mentioned factors highly contributed to the classification. Predominantly-male risk factors included personality, mental health, neurocognitive and brain factors. Conversely, predominantly-female risk factors included environmental factors.

The conclusions of this work showed that multimodal risk factors such as cannabis dependence and use levels demonstrated that environmental factors contributed more strongly to CUD in women, whereas individual factors such as personality and mental health factors had a larger importance in men.

3.3 Parkinson diagnosis based on non-motor symptoms

Martinez-Eguiluz et al. [39] evaluated nine algorithms to be able to differentiate between patients with Parkinson's disease and those in control based on non-motor symptoms only. The algorithms used were AdaBoost, Bagging, Decision Tree, KNN, Multilayer Perceptron, Naive Bayes, Random Forest, Ripper and Support Vector Machine. All this has been carried out thanks to two databases. The first one called Biocruces (which has 96 subjects) and the second PPMI [38] (consisting of 687 subjects). These two databases are made up of attributes which represent different non-motor symptoms of Parkinson's disease. One of the things that was done was an evaluation of whether the combination of the two databases would mean an improvement in the results obtained. In addition

to this, two different versions were created for each database and a feature selection was applied to reduce the number of features, which improved the results obtained. The results obtained show that most of the nine algorithms were able to detect Parkinson's patients as they had achieved an accuracy of more than 80%. The final conclusion of the study demonstrates how PD patients can be detected with high accuracy by applying different machine learning techniques to non-motor symptoms of the disease. It also shows how the most important variables can be selected in order to improve the results.

4. CHAPTER

Theoretical Concepts

4.1 Machine Learning methods

Machine learning is a part of artificial intelligence and computer science that uses data and algorithms to recreate how the humans learn. The main objective is to achieve that a machine can 'learn' in an autonomous way. Machine learning algorithms are generally categorized as supervised or unsupervised [31]. Supervised algorithms are able to apply what has been learned in the past to new data whereas unsupervised algorithms can draw inferences from databases.

The three methods used in this projects are inside the category of supervised learning algorithms and in the following section, theoretical explanations about this methods will be provided.

4.1.1 SVM

Support Vector Machine (SVM) is one of the many algorithms in data mining and also considered as one of the most robust and accurate methods among the well-known data mining algorithms [1]. SVM is also categorized as a new neural network algorithm for forecasting [18]. It is increasingly being used in the areas of research and industry due to its highly effective model in solving non-linear problems [74]. SVM was originally proposed to construct a linear classifier in 1963 by Vapnik [64].

This algorithm is a really powerful method for building a classifier. It aims to create a decision boundary that enables the prediction of labels from one or more feature vectors [46]. This decision boundary is known as the hyperplane. Figure 4.1 shows the classification of a linear SVM.

In a p -dimensional space, a hyperplane is defined as a flat, affine subspace of $p-1$ dimensions. The term affine means that the subspace does not have to pass through the origin. In two-dimensional space, the hyperplane is a 1-dimensional subspace, i.e. a line. In three-dimensional space, a hyperplane is a two-dimensional subspace, a conventional plane. For dimensions $p > 3$, it is not very intuitive to visualize a hyperplane, but the concept of a subspace with $p-1$ dimensions remains. The mathematical definition of a hyperplane is quite simple. In the two-dimensional case, the hyperplane is described according to the equation of a line:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0 \quad (4.1)$$

Given parameters β_0 , β_1 and β_2 , all pairs of values $X = (X_1, X_2)$ for which the equality is satisfied, are points of the hyperplane. This equation can be generalized for p -dimensions:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p = 0 \quad (4.2)$$

and similarly, all points defined by the vector $(X = X_1, X_2, \dots, X_p)$ that satisfy the equation belong to the hyperplane. When X does not satisfy the equation:

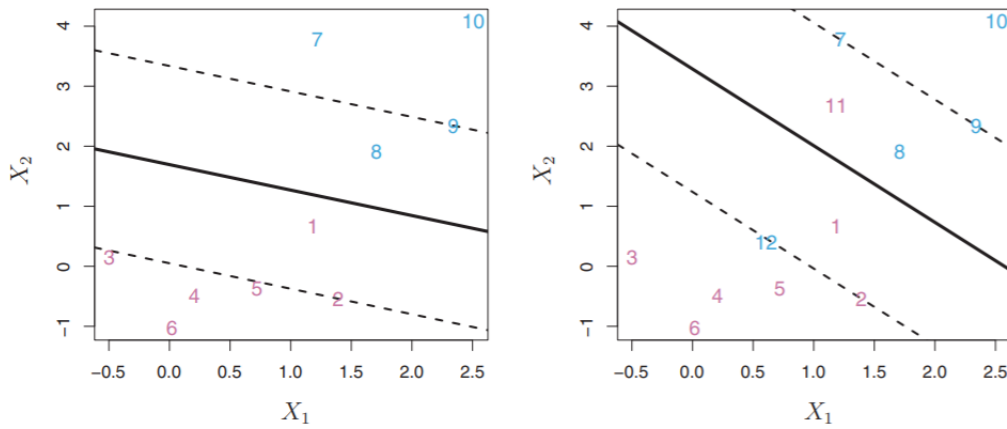
$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p < 0 \quad (4.3)$$

or

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p > 0 \quad (4.4)$$

the point X falls on one side or the other of the hyperplane. Thus, it can be understood that a hyperplane divides a p -dimensional space into two halves. To find out on which side of the hyperplane a given point X lies, one only has to calculate the sign of the equation.

As it can be seen in figure 4.1, in the left image 8 and 1 has crossed the margin (9, 3, 5, 2 are support vectors), whereas on the right image 1 and 8 has crossed the margin (9, 7, 12, 2 are support vectors), and 11 has crossed the hyperplane.



4.1 Figure: SVM hyperplane.

However, there are few cases in the real world where a linear separator can achieve perfect accuracy. This means that it will not be able to solve most of real-world problems accurately. This time, we can achieve non-linearity without modifying the internals of the classifier by applying a kernel function.

For non-linearly separable data problems, the algorithm SVM has been extended using Kernels [11]. Kernels are mathematical functions that transform the data from given space (known as Input Space) to a new high dimensional space (known as Feature Space) where data can be separated with a linear surface (called hyperplane). Figure 4.2, represents the geometrical view of Kernels. Mathematically, a Kernel is a function that takes two arguments, apply a mapping on the arguments and then return the value of their dot product. Suppose x_1 and x_2 are two data points, ϕ is a mapping and K denotes Kernel which is given by:

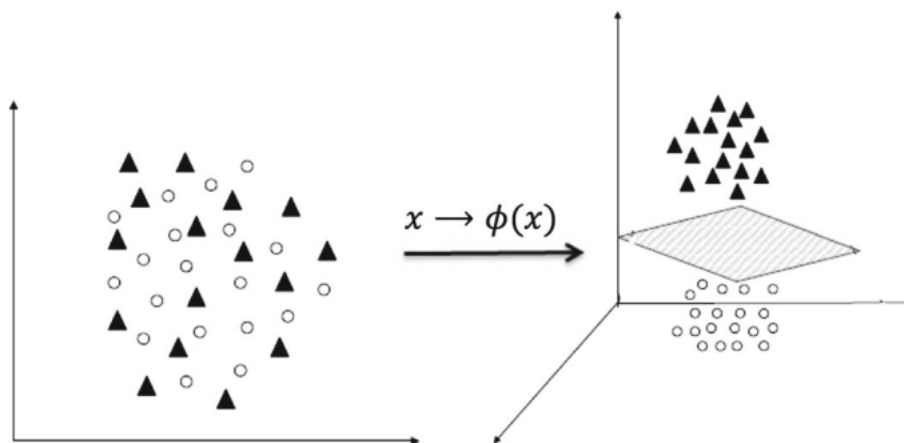
$$K(x_1, x_2) = \phi(x_1)^T \phi(x_2) \quad (4.5)$$

In some applications, the Input Space is rich in features and it is sufficient to take the mapping ϕ as an identity mapping, i.e., $\phi(x) = x$. Kernel where mapping is identity mapping, that is, Input Space and Feature Space are equal, is called linear Kernel and SVM using linear Kernel is called linear SVM. Mathematically, linear Kernel is given by:

$$K(x_1, x_2) = x_1^T x_2 \quad \implies \quad \phi(x) = x \quad (4.6)$$

Linear SVM is very efficient in high dimensional data applications. While their accuracy on test set is close to the non-linear SVM, it is much faster to train for such applications. For example, document classification applications have high dimensional Input Space and there is no need to add more features to the Input Space because it does not make much difference in the performance. So test accuracy of linear SVM is close to that of non-linear SVM [27] but at the same time, training of linear SVM is much faster than non-linear SVM due to the fact of differences between linear and non-linear SVM in their computational complexities.

What can be seen in figure 4.2 is that the left part represents data in Input space and right part represents data in Feature space. Data has been transformed from Input space to Feature space using Kernels. Initially Input space is two dimensional and data is inseparable. Kernels transformed the data to three dimensional space where data is separable by a hyperplane.



4.2 Figure: SVM Kernel.

4.1.2 Multilayer perceptron

Artificial Neural Network

An Artificial Neural Network (ANN) is a mathematical model which has a highly connected structure similar to brain cells [17]. This model consist of a number of neurons arranged in different layers, an input layer, an output layer and one or more hidden layer [17].

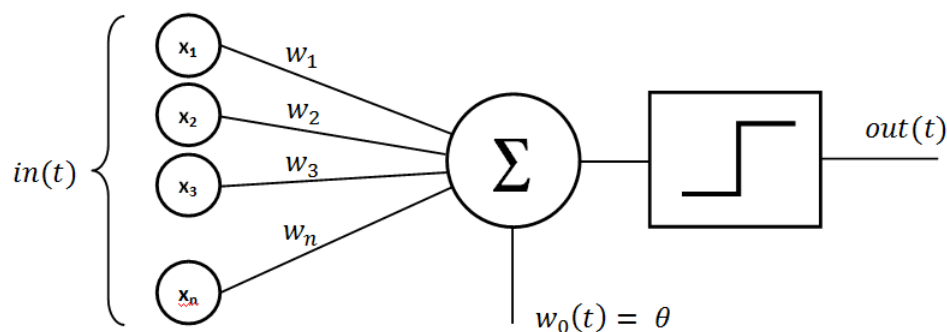
The main purpose while designing an ANN was to emulate human intelligence [48]. The architecture of artificial neural networks is composed of vertices and edges. If we were talking about a biological neural network, instead of vertices and edges, the architecture would be composed of neurons and axons.

One of the greatest contribution of ANNs [30] is that, even if they are capable of imitating a biological neural network, the level of programming that is needed for solving complex problems is very low. These problems are non-analytical, non-linear, non-stationary and even stochastic ones.

Perceptron

The most simple ANNs, which are composed of a single neuron, are called perceptrons. The figure ... shows the architecture of a perceptron. After getting inspiration from the biological neuron and its ability to learn, the perceptron was first introduced by American psychologist, Frank Rosenblatt in 1957 [54].

A perceptron (see figure 4.3) works by taking in some numerical inputs and multiplying these inputs with the respective weights (this is known as the weighted sum) . These products are then added together along with a real-valued bias. After this, a non-linear function is applied to the overall sum. This can be seen in the formula 4.7. This last non-linear function is called activation function and it is used for limiting the output of a neuron [33].



4.3 Figure: Perceptron.

$$\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} [x_1, x_2, \dots, x_n] + b = w_1x_1 + w_2x_2 + \dots + w_nx_n + b * 1 = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \\ b \end{bmatrix} [x_1, x_2, \dots, x_n, 1] \quad (4.7)$$

The most commonly used activation functions are the following ones [58] and they can be seen in figure 4.4 :

- Sigmoid: is the most commonly used activation function and it is a non-linear function. This function transforms the values in the range 0 to 1. It can be defined as:

$$f(x) = 1/e^{-x} \quad (4.8)$$

- Tanh: this is an Hyperbolic Tangent function. Sigmoid function and Tanh function are similar. The main difference is that Tanh is symmetric around the origin. The result of this is that there will be different signs of outputs from previous layers which will be fed as input to the next layer. It can be defined as:

$$f(x) = 2 * sigmoid(2 * x) - 1 \quad (4.9)$$

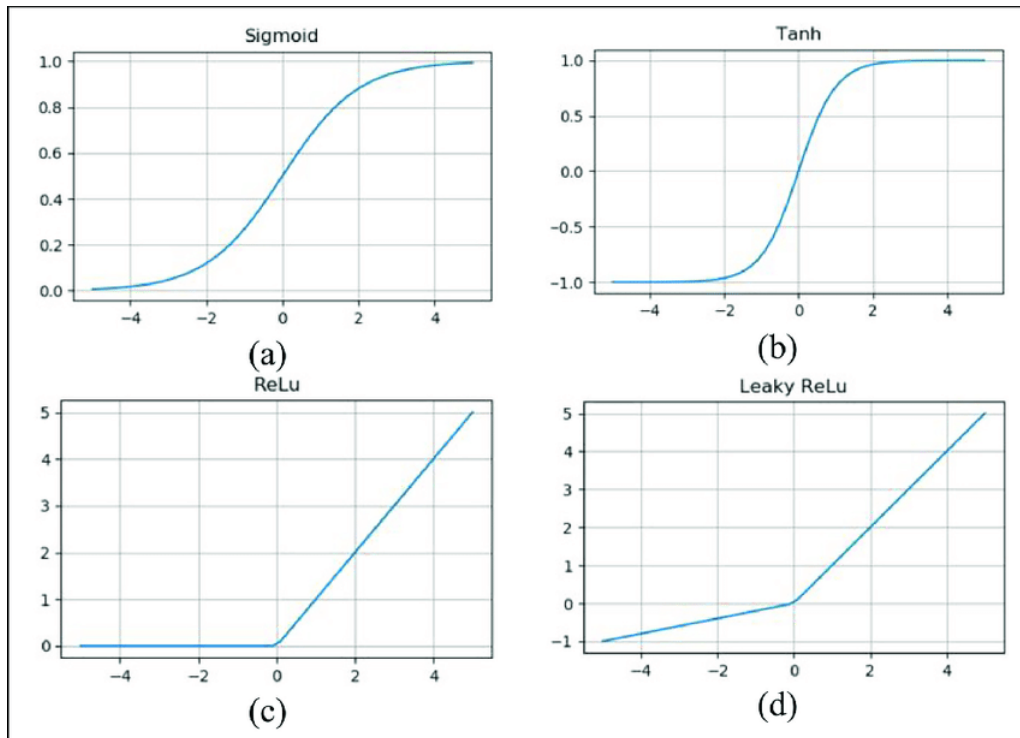
- ReLU: stands for rectified linear unit and is a non-linear activation function which is widely used in neural networks. One of the principles of using ReLU is that all the neurons are not activated at the same time. This means that a neuron will be deactivated only when the output of the linear transformation is zero. It can be defined as:

$$f(x) = max(0, x) \quad (4.10)$$

- Leaky ReLU: is an improved version of ReLU. The main improvement is that for negative values of x, instead of defining the ReLU functions value as zero, it is defined as a really small linear component of x. It can be defined as:

$$f(x) = 0.01x, x < 0 \quad (4.11)$$

$$f(x) = x, x \geq 0 \quad (4.12)$$



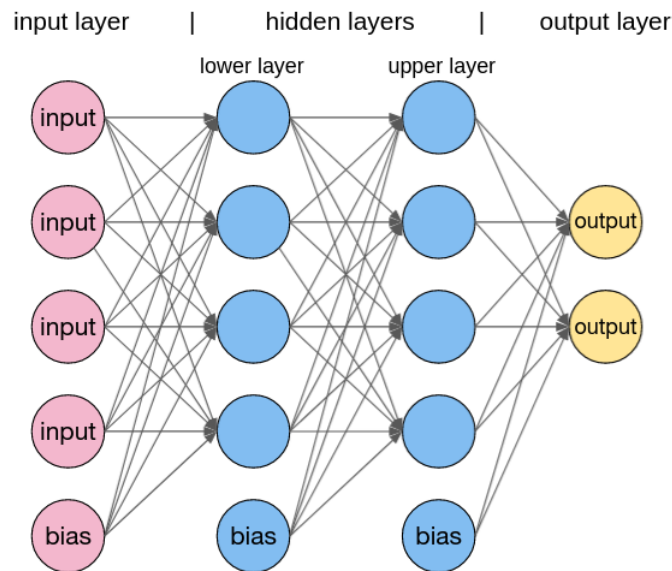
4.4 Figure: Most widely used activation functions.

The same way as linear models are not capable of solving linearly separable problems, perceptrons have the same problem. The reason behind this is because they compute a linear combination of the inputs, even though the result is non-linearized.

Multilayer perceptron

The Multilayer Perceptron (MLP) is a system of simple neurons or nodes interconnected with each other [28]. MLP is a layered feed-forward neural network where the information is flowing through the hidden layers from the input layer to the output layer [7]. These networks are called feed-forward networks because the calculation is sequentially performed layer by layer up to the output layer, although operations in each layer can be done in parallel. A MLP may have one or more hidden layers (the ones that can not be seen by the user) and in the end, an output layer (the last layer of the network). A MLP with two hidden layers can be seen in figure 4.5. The connection between neurons has an

own weight. A really important feature is that MLPs have a non-linear activation function. If not, a single layer perceptron could easily compute it due to the fact that the entire network would be a linear transformation [36]. Another important feature is the need of having differentiable activation functions.



4.5 Figure: Multilayer perceptron.

Loss function

The loss function is the function in charge of computing the existing distance between the output achieved by the algorithm and the expected one. In general words, it's a method for evaluating how the algorithms model the data. It can be categorized in classification (discrete values, 0, 1, 2, ...) and regression (continuous values). The following functions are the most common loss functions:

- Cross Entropy Loss:

$$-\sum_{i=1}^n (y_i \log(z_i) + (1 - y_i) \log(1 - z_i))$$

- Mean Square Error (MSE):

$$\frac{1}{n} \sum_{i=1}^n (y_i - z_i)^2$$

- Mean Absolute Error (MSA):

$$\frac{1}{n} \sum_{i=1}^n |y_i - z_i|$$

Gradient Descent

Gradient descent is one of the most popular algorithms to perform optimization and by far the most common way to optimize neural networks [55]. The way this algorithm calculates the next point is by applying gradient at the current position. After this, thanks to a learning rate η , is scaled and a step is made by subtracting the obtained value from the current position. The reason behind making a subtraction is because we want to minimise the function. If the goal would be to maximize, an addition would be made. This process can be written mathematically as:

$$p_{n+1} = p_n - \eta \nabla f(p_n) \quad (4.13)$$

The parameter which scales the gradient and controls the step size is η . In machine learning terms, it is called learning rate and it has a really important influence in the performance.

The smaller learning rate the longer Gradient Descent converges; or even could reach the maximum number of iterations before the optimum point is reached. If there is the case that the learning rate is too big, there is the possibility for the algorithm not converging to the optimal point or for diverging completely. In summary, the steps for the Gradient Descent method's are:

1. Choose a starting point (initialisation).
2. Calculate gradient at this point.
3. Make a scaled step in the opposite direction to the gradient (objective: minimise).
4. Repeat points 2 and 3 until one of the criteria is met:
 - Maximum number of iterations reached.
 - Step size is smaller than the tolerance.

Backpropagation

The existence of some ANNs with millions of parameters leads to find a proper method for finding the best parameters to converge the loss function. This is a fundamental process. The backpropagation algorithm [28] is the most computationally straightforward algorithm for training the multilayer perceptron. The full mathematical explanations about the algorithm can be found in [7] so only a general overview of the algorithm will be made. In the very first stage, the weights in the network are set to small random values. The next step that the algorithm does is to calculate the local gradient of the error surface and change the weights in the direction of steepest local gradient. When a reasonable smooth error surface is given, the weights could converge to the global minimum of the error surface. In summary, the steps for the backpropagation algorithm are:

1. Initialise network weights.
2. Present first input vector, from training data, to the network.
3. Propagate the input vector through the network to obtain an output.
4. Calculate an error signal by comparing actual output to the desired (target) output.
5. Propagate error signal back through the network.
6. Adjust weights to minimise overall error.
7. Repeat steps 2–7 with next input vector, until overall error is satisfactorily small.

4.1.3 XGBoost

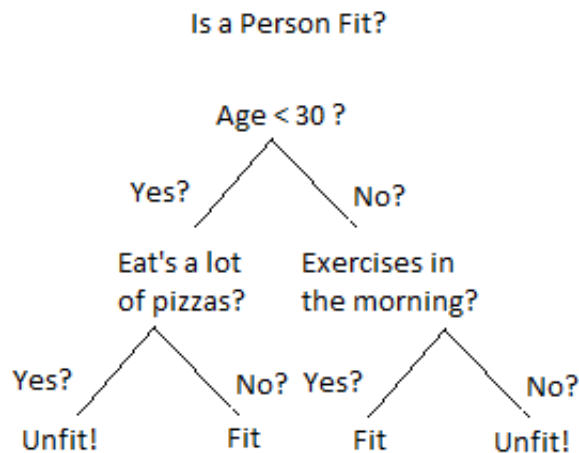
XGBoost is a tree based ensemble machine learning algorithm which is a scalable machine learning system for tree boosting. That is why a little introduction to decision trees and gradient boosting will be made.

Decision Trees

Decision tree (DT) is one of the most powerful and popular tools for classification and prediction [42]. DTs divide data into different subgroups and propose some questions related to the attributes of data items. These questions are contained in a node that points to their child nodes. These child nodes make reference to the possible answers to the questions that can be yes-or-no. A clear example of this concept can be seen in figure 4.6. The way

an item is sorted follows a path that starts in the very first node located on top and finishes in a leaf [35].

On the one hand, one of the main advantage of these models, comparing to other ones as neural networks and support vector machines, is that they are transparent models. This is due to the combination of simple questions about the data. On the other hand, when small changes are applied, huge changes can be noticed in the constructed tree. Unfortunately, small changes in input data can sometimes lead to large changes in the constructed tree. DTs are able to support classification problems with more than two classes and can be modified to handle regression models. When the construction is made, new data is classified in a quick way [35].



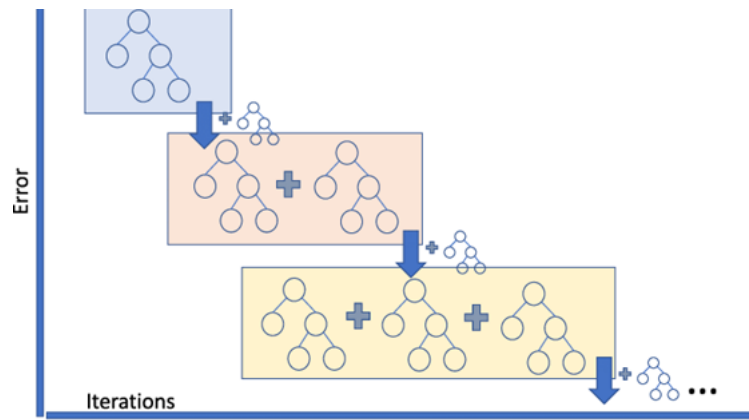
4.6 Figure: Decision Tree classification.

Gradient Boosting

Gradient Boosting (GB) is one of the techniques in the world of machine learning. For many years, it has been the leading technique to achieve state-of-the-art results in different problems such as web search, recommendation systems, weather forecasting and many other [9][53][71][73]. By combining weak models in an iterative way (see figure 4.7) [23][25], boosting algorithms achieve really strong results thanks to a greedy procedure that corresponds to gradient descent in a function space [19].

Decision trees are used as base predictors in most of the popular implementations of GB. Although it is convenient to use decision trees for numerical attributes, most of the data-

bases have a set of categorical attributes (attributes that don't have to be compared with others). Thus, one of the approaches to deal with this problem is to transform the values of those attributes into numbers [4].



4.7 Figure: Gradient Boosting.

XGBoost

XGBoost (XGB) is short for eXtreme Gradient Boosting. It is an efficient and scalable implementation of gradient boosting framework by Friedman et al. (2000) [24].

XGB is a generalised gradient boosting implementation that includes a regularisation term. This term is used to combat overfitting and as a support for arbitrary differentiable loss functions [41].

The main reason of designing XGB was for speed and performance using gradient-boosted decision trees [16]. The main representation of XGB is for machine learning. What it has been taken up nowadays by many developers in the world of machine learning was first done by Tianqi Chen in 2016 [13]. This algorithm is capable of performing the three major gradient boosting techniques: Gradient Boosting [44], Regularized Boosting [8] and Stochastic Boosting [26].

XGB is an algorithm that functions around tree algorithms. One of the main characteristics of tree algorithms is that they take into account the attributes or features of the database. XGB also applies decision-tree algorithms for classifying the data correctly. The main objective is to find the most appropriate parameters from the database used for training purposes. Thus initially, an objective function is set up for describing the performance of the model. This objective function [12] is formed by two different parts. The first one is termed a training loss and the second one is termed a regularization.

$$obj(\theta) = TL(\theta) + R(\theta) \quad (4.14)$$

The parameters are represented by θ , the regularisation term by R and TL makes reference for Training Loss. To measure how predictive the model is TL is used whereas R is helpful to maintain the model's complexity by facing problems such as overfitting. The leading role of XGB is to add the prediction of all trees formed from the database and optimize the result.

However, there is a question that is mainly being asked. How to set up the trees in terms of the parameters used. By calculating the structure of the trees and their respective leaf scores (represented by a function f_t) the parameters can be determined. What XGB tries to do is to optimize the tree that is being learned (training) and at every step taken, adding a tree. To make this, a new objective function (at step t) must be defined. This objective function takes the form of an expansion represented by Taylor's theorem.

$$obj^{(t)} = \sum_{i=1}^n \left[m_i f_t(p_i) + \frac{1}{2} c_i f_t^2(p_i) \right] + R(f_t), \quad (4.15)$$

where m_i and c_i are taken as inputs and p_i refers to the trained data.

The result reflects the desired optimization for the new tree that wants to add itself to the model and this is how XGB deals with loss functions such as logistic regression [16].

Moving on, when defining the complexity of the tree $R(f)$, regularization is really important. The definition of tree $f(o)$ can be defined as [12]:

$$f_t(p) = \omega_{p(p)}, \omega^L, q : R^d \rightarrow 1, 2, 3, 4, \dots, L, \quad (4.16)$$

where ω makes reference to a vector for leaf scores and function q assigns leaves to the corresponding data points. The number of leaves is represented by L . The complexity can be defined as [16]:

$$R(f) = \alpha + \frac{1}{2} \beta \sum_{j=1}^L \omega_j^2. \quad (4.17)$$

For getting the new optimized objective function at step t , the equation of regularization 4.16 can be put in equation 4.15. The model obtained makes a representation of the new reformed tree model and is a measure of how good the tree structure $q(p)$ is.

To establish the tree structure, the following computations must be done at each level: regularization, leaf scores and objective function. As the leaves are split into a right and a left leaf, the gain is calculated at each level at the current leaf with the regularization achieved at the possible additional leaves. If the obtained gain is smaller than the additional regularization value, the branch is abandoned. This is also known as pruning.

After all this mathematical concepts, a general overview of how XGB runs deep into trees and classifies data has been provided; as well as how accuracy and other parameters are calculated.

4.2 Fairness metrics

Recent years have brought extraordinary advances in the field of Artificial Intelligence (AI) [65]. AI now replaces humans at many critical decision points, such as who will get a loan and who will get hired for a job. One might think that these AI algorithms are objective and free from human biases, but that is not the case. It is obvious that accuracy is very important when evaluating the performance of the classification models. However, accuracy and fairness are two concepts that do not go hand in hand. It is said that a tradeoff between accuracy and fairness exists [60].

The existing tradeoff leads to find a balance between these two concepts mentioned before. That is why it is necessary to take a look to some measurements that consider differences between groups. The measurements that we will take a look are demographic parity and bias score. The main motivation to use these two metrics is to check if there is a considerable gap between sex and also between the different values of the class variable.

In order to detect discriminatory outcomes in Machine Learning predictions, we need to compare how well our model treats different user segments. If we calculate the two measurements for the initial sample and the classification, we will be able to know which are the characteristics of the samples and how this changes for each of the generated systems.

4.2.1 Demographic parity

Demographic parity, also known as the lack of disparate impact [61], is the ratio between positives in the minority and majority group. We can write this mathematical expression as:

$$\frac{P(Z = 1|A = 0)}{P(Z = 1|A = 1)}, \quad (4.18)$$

where Z makes reference to the class and A is the sensible attribute which makes reference in this project to the sex. As for instance $A=0$ refers to women and $A=1$ refers to men.

If the achieved value is close to 1 this indicates that the proportion of positives in both groups is similar, achieving demographic parity when the ratio equals 1.

4.2.2 Bias score

The following metric, proposed by [75] evaluates the correlation between a sensitive attribute and the class attribute in our case. Mathematically, we can write it as:

$$\frac{c(Z = 1, A = 0)}{c(Z = 1, A = 0) + c(Z = 1, A = 1)} \quad (4.19)$$

where $c(Z = z, A = a)$ counts the number of appearances of $Z = z$ and $A = a$ altogether. A value closer to 0.5 indicates less bias. One of the differences between demographic parity and bias score is that bias score uses counts instead of probabilities. Thus for instance, in the case of being sex the sensible variable, if the proportion of males and females in the database is not balanced, demographic parity will have a really good value if half of the people in each group are classified with PD. However, bias score will indicate a really high disproportion. In the case that the equality in the number of examples classified between groups is required instead of the equality in the proportion, bias score can be of a great use.

4.3 Evaluation

Systems can't be created and evaluated with the same data. That is why there exists the need of splitting the data to perform the training and the evaluation. One of the techniques that is widely used is cross-validation (CV) [51]. This statistical method helps to compare and select the model in the applied machine learning. The understanding and application of this predictive modeling problem is easy and straightforward. This technique has less bias in estimating the model skills. In our case, the database will be divided into 10 folds. Thus for each $1 \leq i \leq 10$, the i th fold is used for testing and the rest for training. CV

allows us to ensure that the whole database is being tested whilst no sample used for training is evaluated.

4.4 SHAP

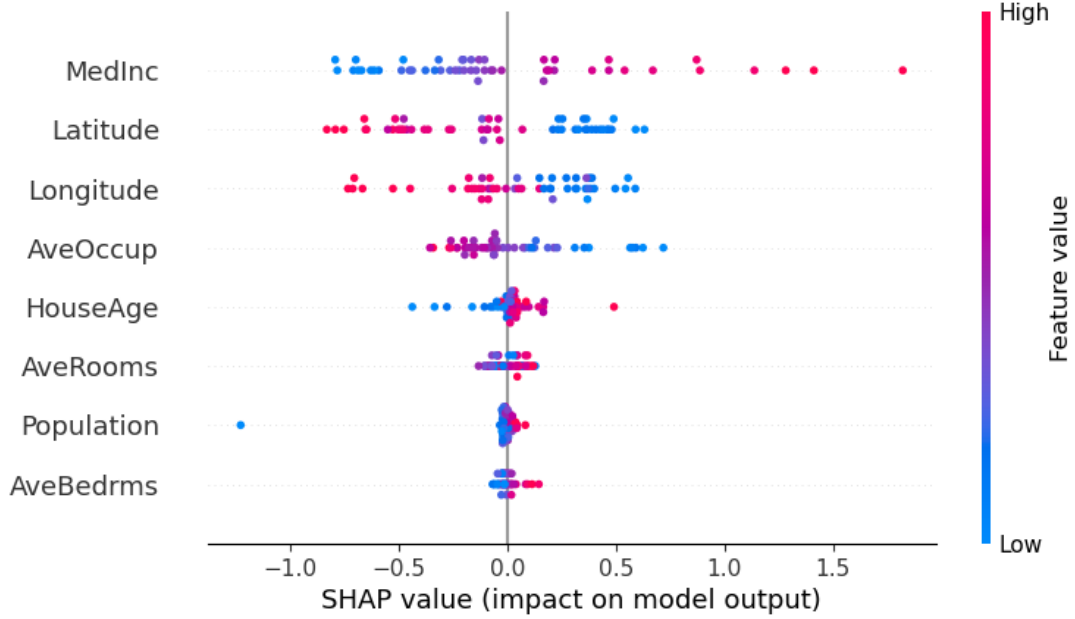
Nowadays, to reason behind a model making a prediction is really important for many different reasons. When trying to understand predictions from tree ensemble methods, importance values are commonly attributed to each input feature. However, the methods in charge of feature attribution for tree ensembles are not really consistent.

To try to tackle this problem a tool called SHAP (SHapley Additive exPlanation) values has been proposed by Lundberg et al. (2018) [37]. The values are the result of combining the ideas from game theory [63] and local explanations [52]. The main objective of this tool is to quantify feature contribution [40]. Using SHAP allows us to attribute a single value, called SHAP value, for each feature of the input for each prediction. This idea can be interpreted as a way of knowing what is the importance of a feature and the influence it has on the prediction.

When interpreting the results obtained with SHAP, high positive or negative values mean that a feature is important. If the values are positive, the feature increases the model's output in terms of the number of classified elements. If the values are negative, the feature decreases the model's output. When the values are close to zero, this suggests that those features have a low importance [69]. To understand this better, figure 4.8 shows an example of how to interpret SHAP values. All variables are shown in an order where the first one is the most important and the last one the less important in terms of global feature importance. It can be seen that high values of Latitude variable have a high negative contribution on the prediction whereas low values have a high positive contribution. The MedInc variable has a really high positive contribution when its values are high, and a low negative contribution on low values. The feature Population has almost no contribution to the prediction, whether its values are high or low.

It has to be taken into account that SHAP shows the contribution or the importance of each feature on the prediction of the model. However, it does not evaluate the quality of the prediction itself.

Lets keep the explanation with some mathematical concepts. To begin with, we will begin with a simple linear regression problem with structured data where the response is continuous [76]. The prediction can be defined as:



4.8 Figure: SHAP values impact on model output.

$$y_i = b_0 + b_1x_{1i} + \dots + b_dx_{di}, \quad (4.20)$$

where we can find that y_i is the i -th prediction response, the corresponding predictors consist of x_{1i}, \dots, x_{di} and b_0, \dots, b_d are the estimated regression coefficients. As mentioned before [37] defined the SHAP value (which is a generalization of the concept explained above this to more complex supervised learning models) to know what was the contribution of the i -th feature as [76]:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \{E[f(X)|X_{S \cup i} = x_{S \cup i}] - E[f(X)|X_S = x_S]\}, \quad (4.21)$$

where F is the whole set of features, S denotes a subset, $S \cup i$ forms the union of the subset S and the feature i and $E[f(X)|X_S = x_S]$ is the conditional expectation of model $f(\cdot)$ when a subset S of features are fixed at the local point x .

5. CHAPTER

Technology

In order to carry out the project properly, it has been necessary to make use of several resources. Among them a development environment, a programming language and some libraries. The resources used will be mentioned below.

GoogleColab

The implementation environment that has been chosen for the realisation of the project has been GoogleColab. Colaboratory, also known as Colab, is a product from Google Research. Colab allows all the users to write and execute arbitrary Python code in the browser. It is ideal for use in machine learning, data analytics and education projects. In technical terms, Colab is a hosted Jupyter notebook service that requires no installation to use and provides free access to computational resources, including GPUs. Colab notebooks are stored in Google Drive, but can also be uploaded from GitHub. Colab notebooks can be shared in the same way as Google Spreadsheets or Google Docs. Another alternative that could have been used was Jupyter. Jupyter is the open source project on which Colab is based. Colab allows you to use Jupyter notebooks and share them with other people without having to download, install or run anything on your computer. The main reason for using Colab is that it has been the most common implementation environment through the whole degree and it provides a lot facilities for executing the code in any device. More information about Colab and how to use it can be find in https://colab.research.google.com/?utm_source=scs-index.

Python

Among all the different existing programming languages, Python has been the one chosen. Python is commonly used to perform data analysis and visualization. It is really easy to learn and has been adopted by many non-programmers such as accountants and scientists, for a variety of everyday tasks. As mentioned before, Python has become a staple in data science, allowing data analysts and other professionals to use the language to conduct complex statistical calculations, create data visualizations, build machine learning algorithms, manipulate and analyze data, and complete other data-related tasks. Python can build a wide range of different data visualizations, like line and bar graphs, pie charts, histograms, and 3D plots. Python also has a number of libraries that enable coders to write programs for data analysis and machine learning more quickly and efficiently, like TensorFlow and Keras. In my case, it should be added that it is the programming language that I have used the most during my studies and the one I am most comfortable with. In addition to this, as the project is focused on data analysis, it is the ideal language. It is also important to take into account the infinity of resources and libraries on the internet about Python. This has facilitated the implementation of the project. Another alternative widely used is R. The problem here is that I do not really like how the interaction with the console works. A tutorial about how to link Python with data science can be found in <https://www.geeksforgeeks.org/data-science-tutorial/>.

Libraries

In the following list the different libraries used for the implementation of the code will be mentioned:

- **Numpy:** NumPy is a Python package that stands for "Numerical Python", it is the core library for scientific computing, providing powerful data structures, implementing arrays and multidimensional arrays. These data structures ensure efficient computations with arrays.
- **Pandas:** allows easy reading and writing of CSV, Excel and SQL database files.
- **Matplotlib:** Matplotlib is a Python library specialised in the creation of two-dimensional graphs. It allows you to create and customise the most common chart types, including the ones used in this project (bar charts).

- Scikit-learn: Scikit-learn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction. The following utilities have been used:
 - >sklearn.model_selection: train_test_split, KFold, cross_val_score, cross_val_predict and MLPClassifier have been possible to use thanks to this library.
 - >sklearn.svm: has allowed to create SVM models.
 - >sklearn.metrics: it has been used for computing different metrics such as confusion_matrix, accuracy_score and classification_report.
- xgb: is the library in charge of the creation of XGBoost models.
- shap: it has allowed all the tasks related with SHAP.

6. CHAPTER

Project development

6.1 Methodology

The project is based on the three papers found as references and will attempt to answer questions such as whether sex is a relevant attribute for classification and whether within PD or HC patients sex can be differentiated. The first step for the realisation of this project has been to analyse the provided databases [39]. Within this step it has been possible to compute statistics and the different fairness metrics such as Demographic parity and Bias score. The next step was the construction of different classifiers for all data, for women only and for men only, where all the evaluations have been performed using 10 fold-CV.. These classifiers have been SVM, MLP and XGBoost. The reason for having made use of both SVM and MLP was because both classifiers obtain the best results in the groundwork [39]. While the decision to use XGBoost was made on the basis that it is the common classifier used in all the reference works [67][45]. Then the analysis of the importance of the attributes has been carried out using XGBoost. The next step was to perform the classification according to sex and at the same time to re-analyse the importance of the attributes using XGBoost once again [67][29]. Finally, the SHAP tool was used to assess the importance of different features and specifically sex in the different proposals [29][45].

6.2 Databases

In this section, a brief explanation will be given about the databases we have worked with. It is worth mentioning that although in the first phases of the project we worked with the database known as Biocruces, it was finally decided to dispense with it. The main reason for this decision was the small number of cases available. The total number of patients is 96, i.e. not even a hundred, and the results obtained were not very convincing. Despite this, as it was part of the project, an explanation is also given about it. A large part of these explanations has been made possible thanks to the work of Martinez-Eguiluz et al. [39].

6.2.1 Biocruces

The Biocruces database consisted of demographic data and a collection of non-motor clinical outcomes from several questionnaires and psychophysical tests. Participants were recruited and evaluated between the years 2015 and 2018 through the Department of Neurology at Cruces University Hospital and the Biscay PD Association (ASPARBI). This database contained information from 96 patients. Within this number of patients, 59 of them with Parkinson (PD) and 37 of them healthy controls (HC). The number of women is 38 compared to 58 men. Among the women, 21 of them have PD and 17 are controls. Among the men, 38 men with PD and 20 controls. This database has 13 attributes and the column associated to the class variable. No use has been made of this database as mentioned in the introduction to this section.

	PD	HC	Total
Women	21	17	38
Men	38	20	58
Total	59	37	96

6.1 Table: Distribution of BIO

6.2.2 PPMI

Parkinson's Progression Markers Initiative (PPMI) [38] is a landmark, multicenter, longitudinal study that aims to identify biomarkers for the progression of PD to improve therapeutic and etiological research.

Only non-motor symptoms have been used in this project. A brief explanation about the different tests and questionnaires that have been taken into account will be added:

- Geriatric Depression Scale (GDS) [72]: is a questionnaire to make reference of the degree of depression in older adults. This questionnaire has 15 questions and depending on the score the results suggest normal (0-4), mild (5-8), moderate (9-11) or severe (12-15) depression,
- Symbol Digit Modalities Test (SDMT) [62]: a test used to detect cognitive impairment. It consists of substituting digits for abstract symbols using a reference key.

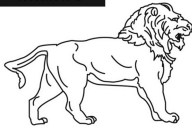
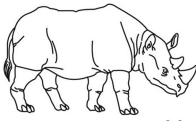
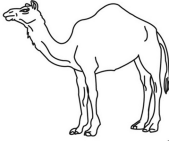
⊂	⊄	⊆	⊇	⊈	⊉	+	⊃	÷
1	2	3	4	5	6	7	8	9

⊂	⊄	⊆	⊇	⊈	⊉	+	⊃	÷
4	5	8	2	7	3	1	9	6

6.1 Figure: SDMT.

- Benton Judgement of Line Orientation Test (BJLOT) [5]: cognitive test of 30 items used to measure a person's ability to match the angle and orientation of lines in space.
- University of Pennsylvania Smell Identification Test (UPSIT) [21]: is a test which has 40 items and it is used for clinically quantifying olfactory deficits.
- Montreal Cognitive Assessment (MoCA) [43]: it evaluates the general cognitive abilities of the subject. MoCA is a neuropsychological test administered by professionals and whose score has a range between 0 and 30.
- Hopkins Verbal Learning Test (HVLT) [3]: the verbal learning and memory of the subject is tested. The test consists of three trials of free-recall of a semantically categorized 12-item list, followed by yes/no recognition.

For this project, two different levels of database were created. The first level makes reference to a database where it takes into account the total scores of each questionnaire and test (feature per test, FPT) and the second one includes the individual items of each questionnaire and test (feature per item, FPI).

NAMING																													
			[]	[]	[]																								
MEMORY	Read list of words, subject must repeat them. Do 2 trials, even if 1st trial is successful. Do a recall after 5 minutes.	FACE	VELVET	CHURCH	DAISY	RED																							
	1st trial																												
	2nd trial																												
ATTENTION	Read list of digits (1 digit/sec.).	Subject has to repeat them in the forward order	[]	2	1	8	5	4																					
		Subject has to repeat them in the backward order	[]	7	4	2																							
Read list of letters. The subject must tap with his hand at each letter A. No points if ≥ 2 errors																													
	[]	F	B	A	C	M	N	A	A	J	K	L	B	A	F	A	K	D	E	A	A	J	A	M	O	F	A	A	B
Serial 7 subtraction starting at 100																													
	[]	93	[]	86	[]	79	[]	72	[]	65																			
4 or 5 correct subtractions: 3 pts. 2 or 3 correct: 2 pts. 1 correct: 1 pt. 0 correct: 0 pt																													

6.2 Figure: MoCA.

6.2.3 FPT

This database contained information from 687 patients. Within this number of patients, 490 of them with PD and 322 of them controls. The number of women is 239 compared to 448 men. Among the women, 168 of them have PD and 71 are controls. Among the men, 322 men with PD and 126 controls. This database has 14 attributes and the column associated to the class variable. Demographic and general clinical attributes included: sex (GENDER), years of education (EDUCYRS), dominant hand (HANDED), age (AGE) and the class itself (PD or control). On the other hand, non-motor data included information from the mentioned neuropsychological tests (SDMT, BJLOT, MoCA and HVLT), from an olfaction test (UPSIT), from a questionnaire on neuropsychiatric symptoms (GDS) and autonomic manifestations (SCOPA-AUT).

	PD	HC	Total
Women	168	71	239
Men	322	126	448
Total	490	197	687

6.2 Table: Distribution of FPT

FPTWomen

Same database as PPMI but where all patients are women. This database contained information of 239 women. Among them, 168 with PD and 71 are controls. The existence of class imbalance is notable.

	PD	HC
Women	82.4%	17.6%

6.3 Table: Distribution of FPTWomen

FPTMen

Same database as PPMI but where all patients are men. This database contained information of 448 men. Among them, 322 with PD and 126 are controls.

	PD	HC
Men	71.9%	28.1%

6.4 Table: Distribution of FPTMen

6.2.4 FPI

This database is almost identical to the one explained above called FPT. The only difference is in the number of attributes. While FPT makes use of 14 attributes and the column associated to the class variable, FPI consists of 97 attributes and the column associated to the class. The reason why the number of attributes is higher is because the individual responses are added for each test; rather than an overall result for each test.

FPIWomen

Same database as FPI but where all patients are women. This database contained information of 239 women. Among them, 168 with PD and 71 are controls.

FPIMen

Same database as FPI but where all patients are men. This database contained information of 448 men. Among them, 322 with PD and 126 are controls.

6.2.5 FPTSex

FPTSex is the same database as FPT with the only peculiarity that the class column instead of being PD or HC, is now the sex (men or women).

	Women	Men	Total
PD	168	322	490
HC	71	126	197
Total	239	448	687

6.5 Table: Distribution of FPTSex

FPTSexPD

Same database as FPTSex but where all patients suffer from PD. This database contained information of 490 patients with PD. Among them, 168 women and 322 men.

	Women	Men
PD	34.3%	65.7%

6.6 Table: Distribution of FPTSexPD

FPTSexHC

Same database as FPTSex but where all patients are HC. This database contained information of 197 patients with PD. Among them, 71 women and 126 men.

	Women	Men
HC	36%	64%

6.7 Table: Distribution of FPTSexHC

6.2.6 FPISex

The same is true for this database. The only difference with FPI is that the class variable is now the sex. Everything else is exactly the same as FPI.

FPISexPD

Same database as FPISex but where all patients suffer from PD. This database contained information of 490 patients with PD. Among them, 168 women and 322 men.

FPISexHC

Same database as FPISex but where all patients are HC. This database contained information of 197 patients with PD. Among them, 71 women and 126 men.

6.3 Results

The aim of this chapter is to provide an analysis about the different results obtained through the different levels of the databases. First of all we will take into account the study made about the classification of Parkinson Disease and discuss the general results of the databases FPT and FPI. Then we will analyse some confusion matrices to determine the precedence of the error and we will continue with the analysis of fairness metrics. We will finish with an interpretation of the graphics related to the feature importance. The same analysis will be made when we take into account the study made about sex classification. Although due to time and space limitations, performance analysis will be mainly based on accuracy. Other metrics such as f-measure, recall and precision have been obtained and are included in appendix A. The values shown in the graphics are average values obtained in a 10 fold-CV.

6.3.1 PD/HC classification

Figure 6.4 and 6.5 show the accuracies obtained by three different classification models (SVM, MLP and XGBoost) in the two different database levels. In the vertical axis, the accuracies obtained are shown whereas in the horizontal axis we can find three subgroups which make reference to each of the classifiers. Each of these subgroups contain the accuracy obtained for different groups of data.

- The first column shows the results obtained with the whole database (FPT/FPI) when differentiating PD from HC. We will call this option global classifier.

- The second and third columns represent the computed accuracies taking into account two different samples for test. The second column has been created by only taking women from the original sample whereas the third columns taking only men.
- The last columns show the results obtained when creating classifiers with databases that contains women and men.

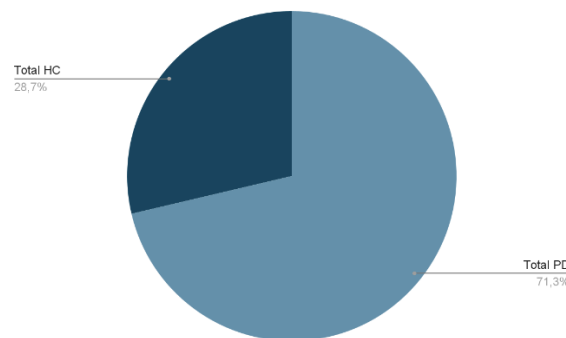
Figure 6.6 and 6.7 make reference to the confusion matrices obtained with the XGBoost classifier for the different classification options described above. The vertical axis refers to the “true label” and the horizontal one to the “predicted label” (where 0 is HC and 1 is PD).

Figure 6.8 and 6.9 are the ones related with the fairness metrics applied in this project (demographic parity and bias score).

The importance of the features are shown in Figure 6.10 using XGBoost plots and in Figure 6.11 and 6.12 using SHAP.

A more detailed analysis for every of the graphics will be made in the following subsections.

Before displaying the results, the distribution of the PD/HC class is shown for easier interpretation.

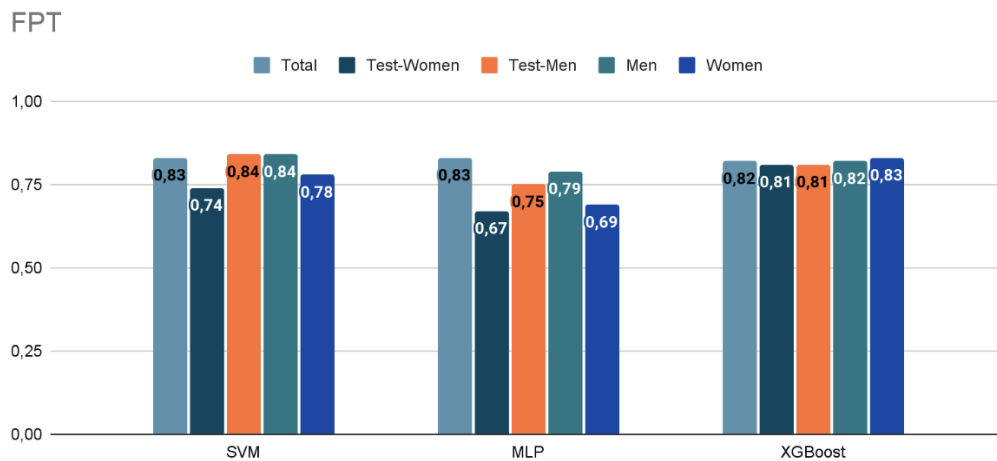


6.3 Figure: PD/HC class distribution.

Accuracy

Regarding accuracy, in the general case, the accuracies obtained in FPT (see figure 6.4) are 0.83, 0.83 and 0.82. The three classifiers (SVM, MLP and XGBoost) are able to pre-

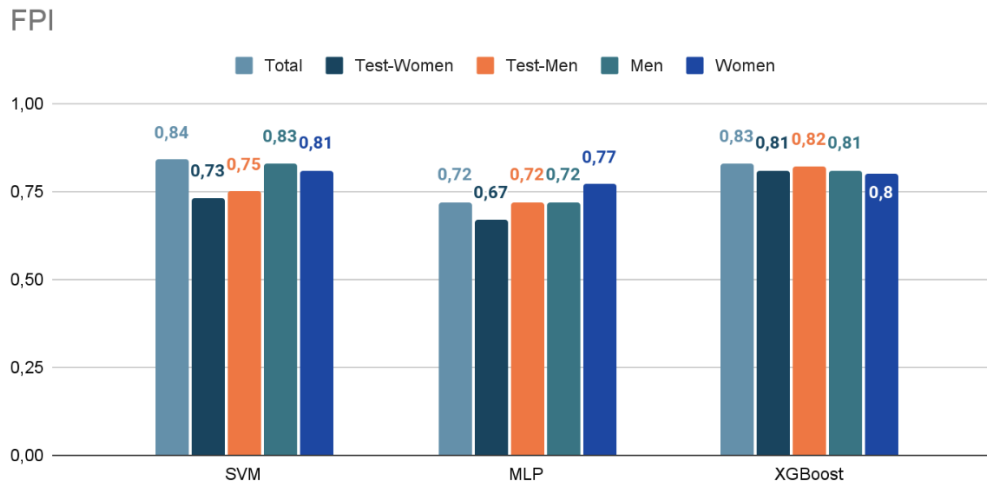
dict PD and HC subjects with more than 80% of accuracy. When making a comparison differentiating the sex, different results can be seen. Inside FPT, when 2 different samples are created, the results obtained are 0.84, 0.75 and 0.81 for the sample created with men and 0.74, 0.67 and 0.81 for the sample created with women. For the database created only with male subjects (FPTMen), the achieved accuracies are 0.84, 0.79 and 0.82; whereas in the database with only female subjects (FPTWomen) the scores are 0.78, 0.69 and 0.83. It can be noticed that when the sex is differentiated, all of the three classifiers behave better in terms of accuracy when they are constructed in the databases where only a unique sex exists (FPTMen and FPTWomen).



6.4 Figure: Accuracies for FPT.

We can make the same analysis in FPI (see figure 6.5). The general accuracies achieved here are 0.84, 0.72 and 0.83. In this case, not all the classifiers obtain an accuracy higher than 80%. However, the results are acceptable and they seem to classify PD and HC subjects correctly. In terms of sex differentiation, the 2 test samples created inside FPI (X_testMen and X_testWomen) show an accuracy of 0.75, 0.72 and 0.82 in the case of men and 0.73, 0.67 and 0.81 in the case of women. Looking into FPIMen the scores are 0.83, 0.72 and 0.81 and in FPIWomen these scores are 0.81, 0.72 and 0.80. Once again, in general terms, the classifiers obtain better performance in the databases containing only men or women. Although it can be noticed that XGBoost behaves a little better in both of the different test samples (X_testMen and X_testWomen).

The results also show a better performance of the classifiers in FPT rather than in FPI.

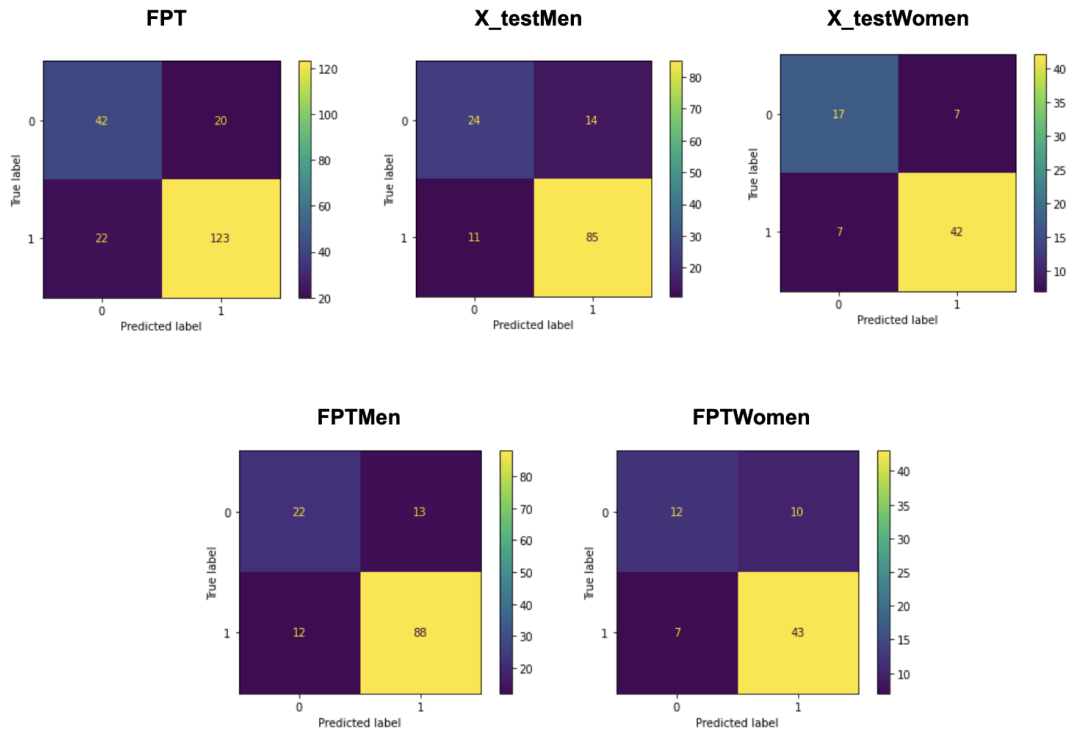


6.5 Figure: Accuracies for FPI.

Confusion matrix

Looking at the confusion matrices, in FPT 123 cases of PD are correctly classified and 22 incorrectly; whereas 42 cases of HC are correctly classified and 20 incorrectly. When differentiating the sex, when testing with men, for PD classifies 85 subjects right and 11 wrong and in the case of HC, 24 right and 14 wrong. The results obtained when testing with women are 42 correct classifications for PD and 7 wrong and 17 correct classifications for HC and 7 wrong. Moving on to the databases created only with men or women, FPTMen has a confusion matrix of 88 correct and 12 incorrect PD classifications and 22 correct and 13 incorrect HC classifications. FPTWomen makes 43 correct and 7 incorrect PD classifications and 12 correct and 10 incorrect HC classifications. If we compare the performance of the classifiers when sex is differentiated, in databases where a unique sex exists (FPTMen and FPTWomen) PD subjects are better classified whereas HC subjects are better classified inside the global classifier. All these results can be seen in figure 6.6.

Making the same analysis in FPI, the correct and incorrect classifications for PD are 127 and 18 whereas for HC are 45 and 17. In the analysis made using different test samples in the global classifier, X_testMen has 85 correct and 11 incorrect classifications for PD and 25 correct and 13 incorrect for HC. Alternatively, for X_testWomen, the correct and incorrect classifications for PD are 41 and 18 whereas for HC are 18 and 6. The PD classification in FPIMen are 90 correct and 10 incorrect and for HC 20 correct and 15 incorrect. In FPIWomen, PD has 45 right and 5 wrong classifications whereas HC has 13 correct and 9 incorrect. So the same way it happened with FPT classifiers, when the



6.6 Figure: Confusion matrices for FPT.

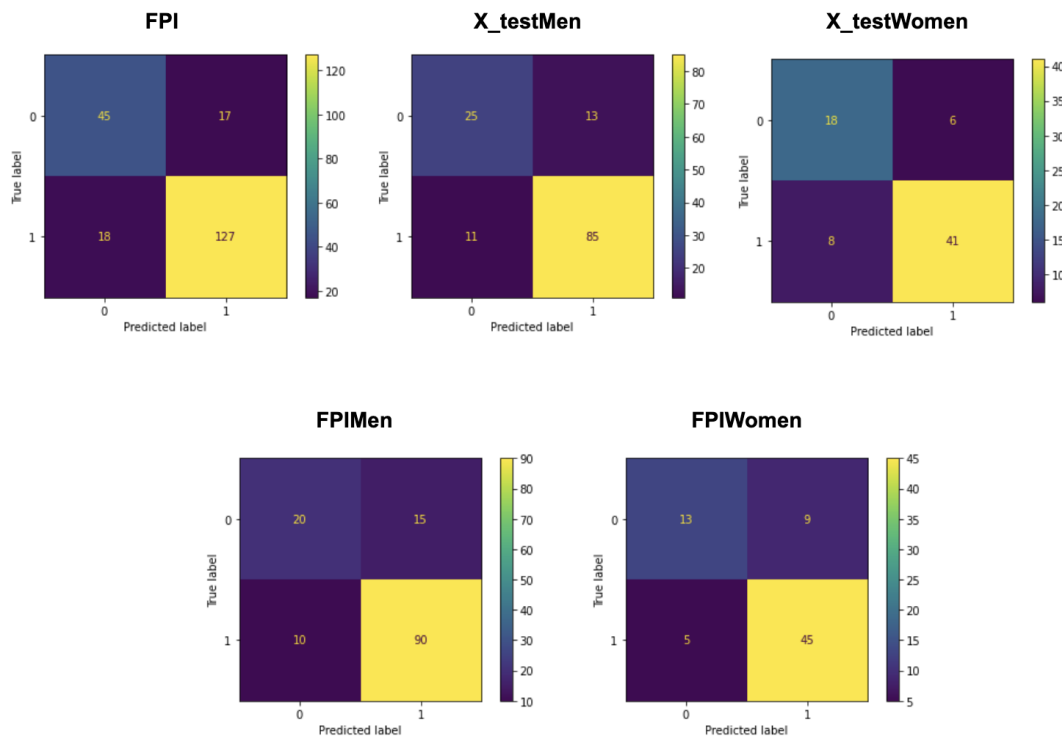
sex is differentiated, the performance of the classification for PD shows a better behavior in FPTIMen and FPTIWomen whereas HC is better classified in the two samples, once again. All the results are shown in figure 6.7.

In this case, the results show a better performance of the classifiers in FPI rather than in FPT. However, if we put the focus on proportions, HC is always not really well classified whereas PD classification is acceptable. This could happen because the number of PD subjects in all the databases is much higher than the number of HC subjects.

Fairness metrics

The metrics of “Demographic Parity” and “Bias Towards PD” have been implemented before performing classification tasks and after doing the classification with XGBoost. It has been decided to evaluate in the whole database and only using XGBoost.

From a practical point of view, maintaining these two metrics in the same level they have in the training sample would mean keeping the balances in the same proportion found in the training sample, and thus not generating or amplifying biases. Regarding



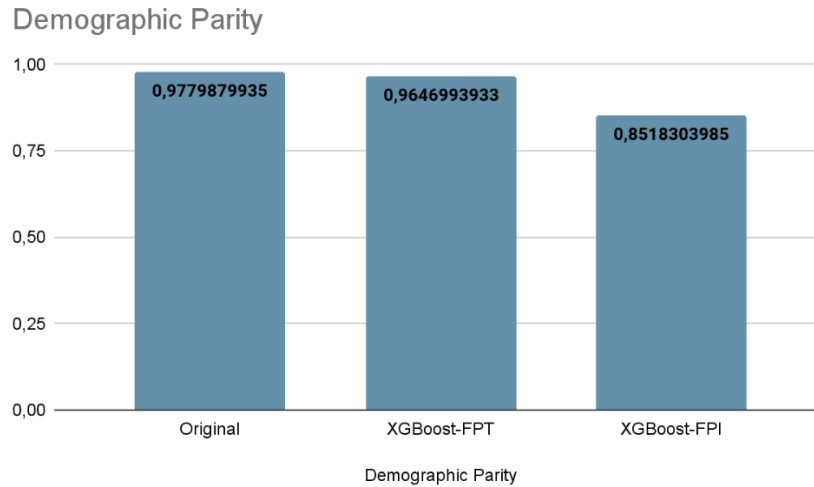
6.7 Figure: Confusion matrices FPI.

demographic parity (see figure 6.8), the results achieved in FPI generate more bias than the ones achieved before doing classification tasks and XGboost in FPT. However, all of the three values are quite close to 1 so that is a good result.

Moving on to bias towards PD (see figure 6.9), the ideal result is 0.5. In this case, the opposite happens. The one that changes the most is FPI.

Feature importance

When we analyze FPT, FPTMen and FPTWomen (see figure 6.10), it can be seen how the most important features are always quite similar. In the graphics obtained by XGboost, the ranking of the features is nearly always the same (“AGE”, “SDMT”, “UPSIT” and “SCAU”). When SHAP is used with the classifiers SVM and XGBoost (see figure 6.11), the same happens. “UPSIT”, “SCAU”, “AGE” and “SDMT” are normally the features with the most importance. In the analysis of FPI, FPIMen and FPIWomen (see figure 6.10), something similar happens. The most important features according to the XGBoost graphics are always “AGE”, “SDMT”, “UPSITBK1” and “MCAVFNUM”. When using

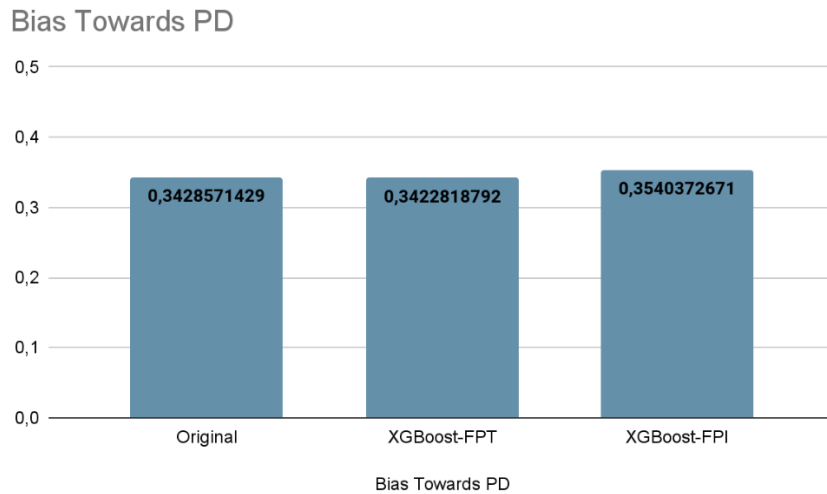


6.8 Figure: Demographic parity for PD/HC classification.

SHAP with the classifiers SVM and XGBoost (see figure 6.12), the features with the most importance are normally “UPSITBK4”, “UPSITBK1”, “AGE” and “SCAU2”. However, in this case, the differences in the ranking of the features between SVM and XGBoost change a little bit more.

As we have used two different methods to analyze the importance of the features, it is convenient to know if they take into consideration the same features. Inside FPT, FPTMen and FPTWomen, both XGBoost and SHAP provide a similar ranking of features. FPI, FPIMen and FPIWomen follow the same pattern of feature ranking as well. However, it is slightly different to the databases generated with FPT. These results make sense because the number of existing features in FPI databases is enormous compared to the ones in FPT and this leads to more diversity of ranking. But it can be noticed that “UPSITBK4”, “UPSITBK1”, “AGE” and “SCAU2” are the ones that appear the most in all the graphics.

As one of the main objectives of the project, it is important to know whether sex plays an important role in the classification tasks. In all the rankings made with XGboost, inside FPT, sex has almost no importance as it is used very few times. In FPI it does not even appear in the top 10 features. When using SHAP with the classifiers SVM and XGBoost, all the graphics computed show that the feature “GENDER” in the model output is almost 0.



6.9 Figure: Bias towards PD.

6.3.2 Sex classification

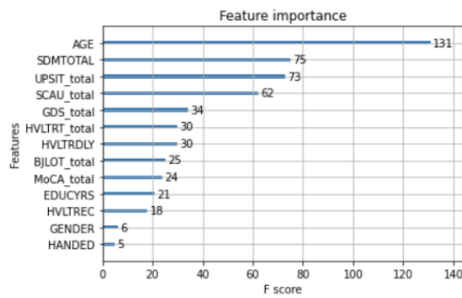
Figure 6.14 and 6.15 show the accuracies obtained by three different classification models (SVM, MLP and XGBoost) in the two different database levels. In the vertical axis, the accuracies obtained are shown whereas in the horizontal axis we can find three subgroups which make reference to each of the classifiers. Each of these subgroups contain the accuracy obtained for different partitions. Each column of these subgroups represent:

- The first column takes into account the results obtained with the whole database (FPTSex/FPISex).
- The second and third columns represent the computed accuracies taking into account two different test samples. The first one has been created by only taking PD subjects from the original test sample and the second one taking only HC subjects.
- The last two columns show the results obtained when creating the classifiers with separated databases for PD and HC subjects.

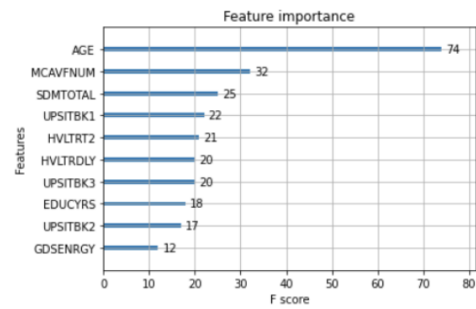
Figure 6.16 and 6.17 make reference to the confusion matrices obtained with the XGBoost classifier for the different approaches. The vertical axis refers to the “true label” and the horizontal one to the “predicted label” (where 0 is women and 1 is men).

Figure 6.18 and 6.19 are the ones related with the fairness metrics applied in this project (demographic parity and bias score).

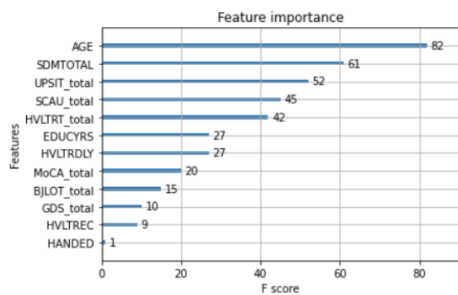
FPT



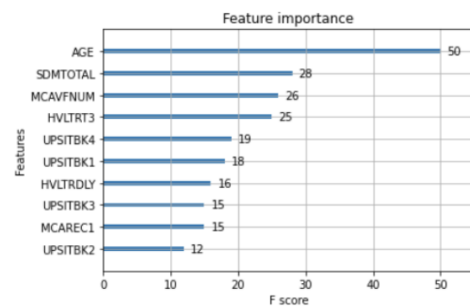
FPI



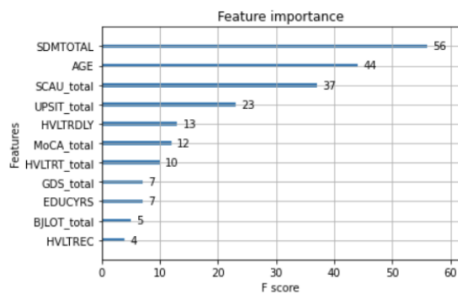
FPTMen



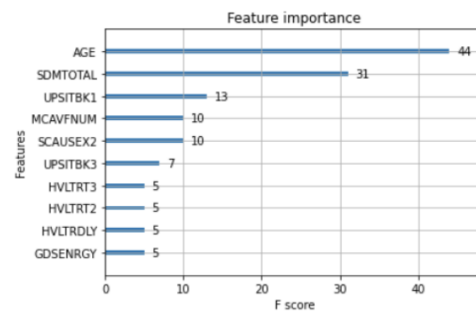
FPIMen



FPTWomen



FPIWomen

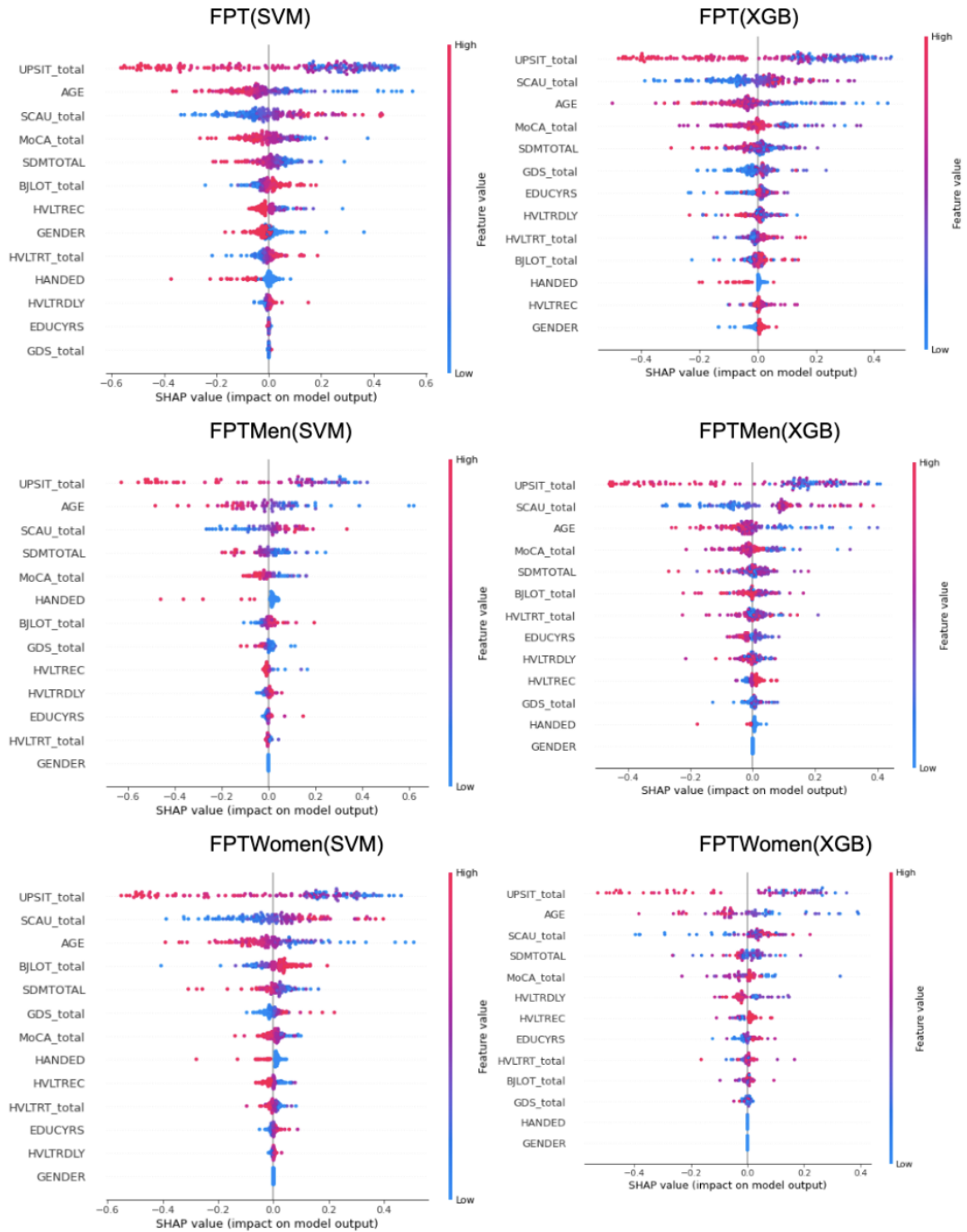


6.10 Figure: XGBoost ranking for PD/HC classification.

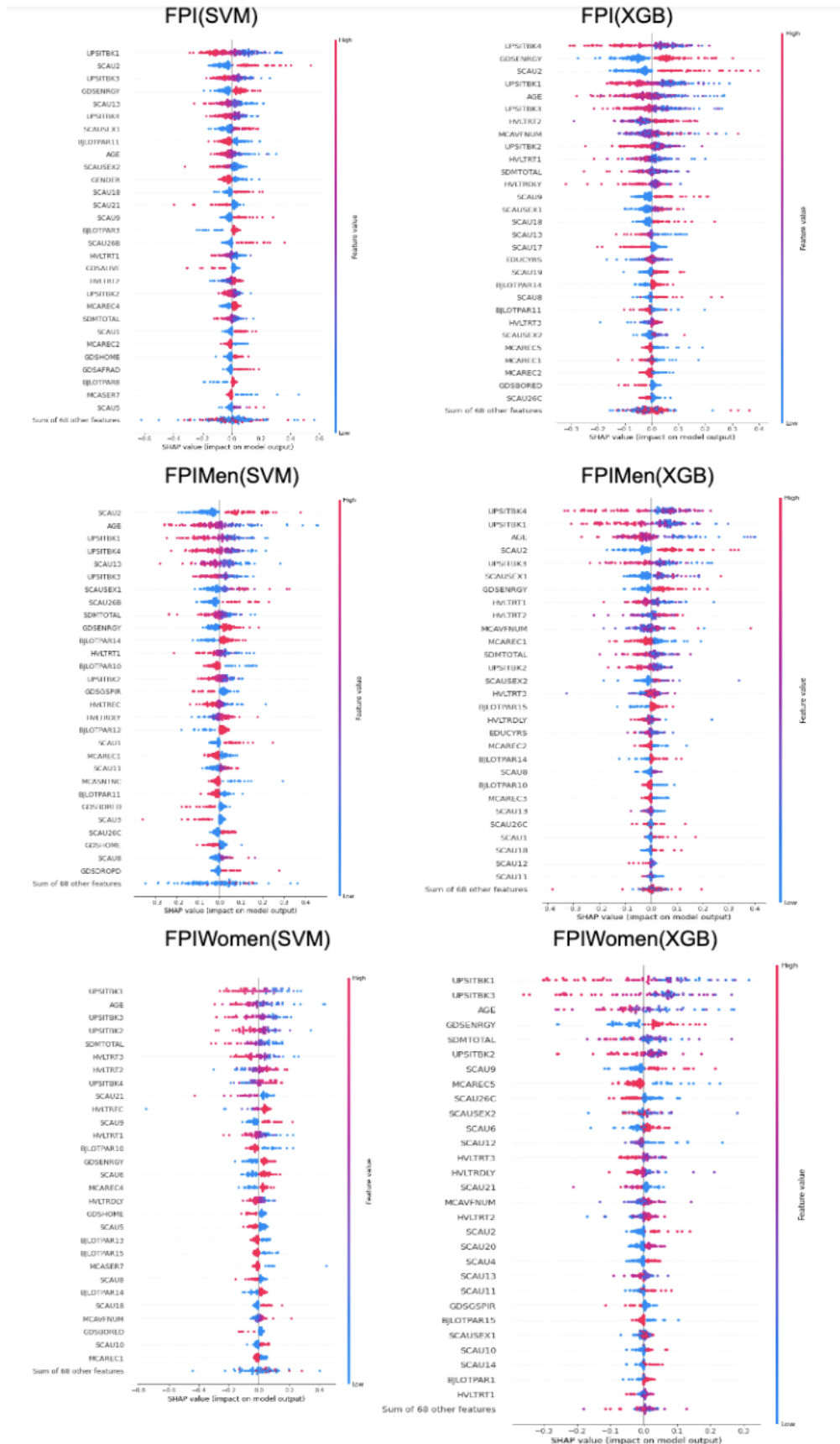
The importance of the features is shown in figure 6.20 using XGBoost.

A more detailed analysis for every of the graphics will be made in the following subsections.

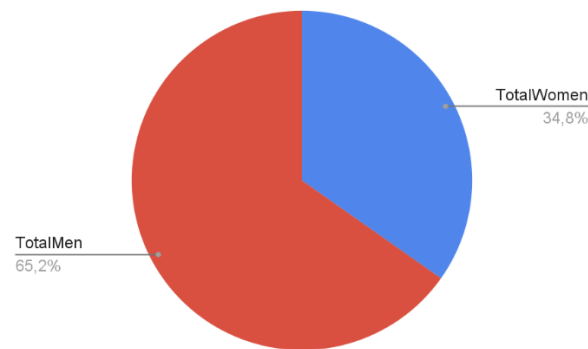
Before displaying the results, the distribution of the sex class is shown for easier interpretation.



6.11 Figure: SHAP in FPT.



6.12 Figure: SHAP in FPI.



6.13 Figure: Sex class distribution.

Accuracy

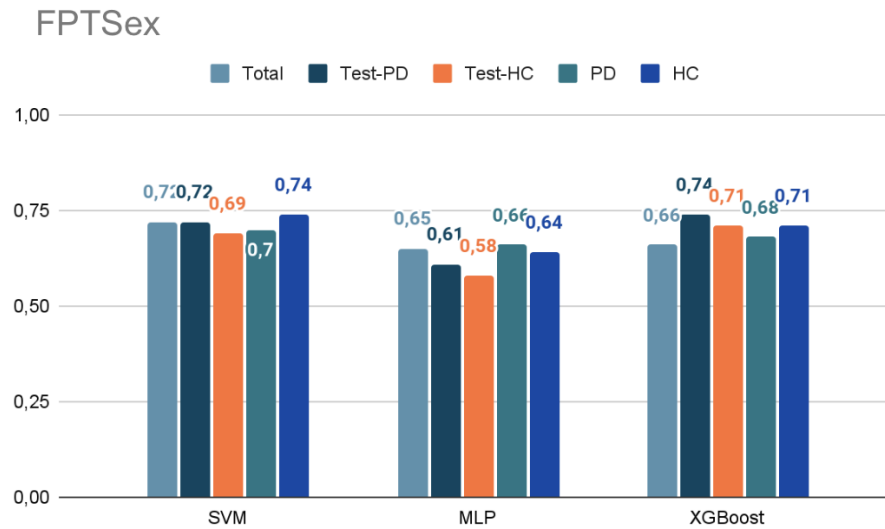
This analysis is going to be divided in two. The first result is the one achieved with SVM, the second one with MLP and the last one with XGBoost. The first analysis will be made among those subjects that suffer from PD and the second one among HC subjects. Regarding the databases created only from PD subjects, in FPTSexPD the accuracies obtained are 0.7, 0.66 and 0.68 whereas in FPISexPD these accuracies are 0.78, 0.77 and 0.80.

Moving to those databases where only HC subjects exist, in FPTSexHC 0.74, 0.64 and 0.71 are the accuracies computed and in FPISexHC 0.74, 0.73 and 0.75.

In general terms, the classifiers in FPISexPD and FPISexHC make a better performance and it seems that the accuracy obtained is better when we have databases where only PD subjects exist. All the results can be seen in figures 6.14 and 6.15.

Confusion matrix

We will analyze the behavior of the confusion matrices to see if the classification of sex is well performed. This analysis is going to be divided in the same way accuracy was divided. The first analysis will be made among those subjects that suffer from PD. In FPTSexPD, 74 men are well classified and 22 bad whereas 24 women are good and 27 bad classified. It can be seen that the classification of women is not well performed. The results in FPISexPD are way better than the ones achieved in FPTSexPD but still not reliables. Men classification is not that bad because 85 are correctly classified and 11



6.14 Figure: Demographic parity for PD/HC classification.

incorrectly classified. Inside women, 30 good and 21 bad classified. The results can be seen in figure 6.16.

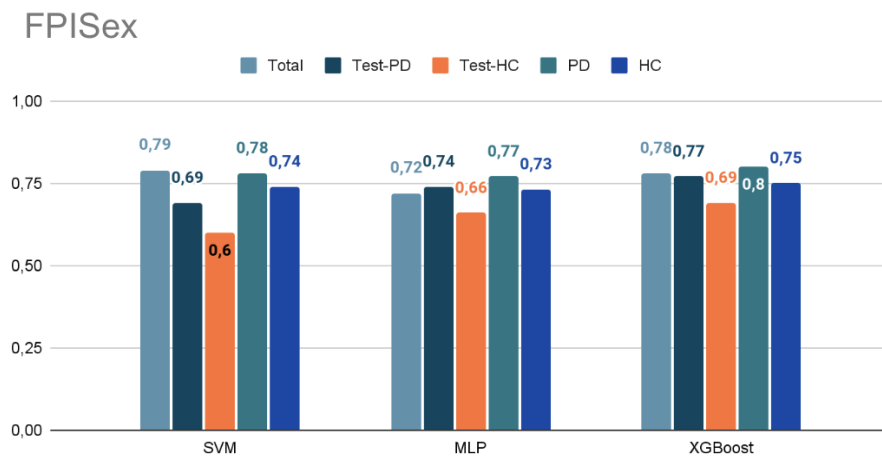
Moving on to the databases formed only with HC subjects, in FPTSexHC 31 men are correctly classified and 10 incorrectly classified; whereas 10 women are good and 9 badly classified. Looking at the results obtained in FPISexHC they are slightly better than the ones obtained before. But still not good enough. Inside men the correct and incorrect classifications are 37 and 4 and inside women 9 and 10. The proportions are quite bad. The results are shown in figure 6.17.

According to these results, it seems that sex cannot be reliably classified according to the data obtained and the tests carried out.

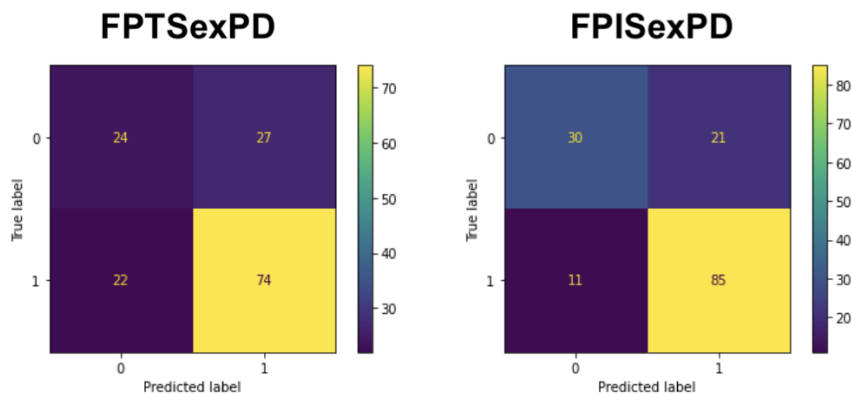
Fairness metrics

The same analysis as the one made with PD/HC will be made here.

Regarding demographic parity (see figure 6.18), the results achieved in FPTSex were worse than the other two scores. However, all of the three values are quite close to 1 and overall the results are better than having PD/HC as the class. Moving on to bias towards uniform class (see figure 6.19), the ideal result is 0.5. Once again the best results are achieved using XGBoost in the FPISex database.



6.15 Figure: Demographic parity for PD/HC classification.

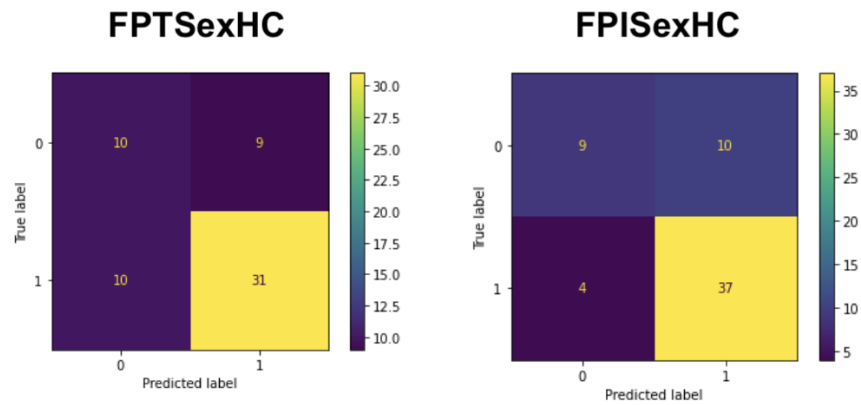


6.16 Figure: Confusion matrices for PD.

Feature importance

In this case, only an analysis using XGBoost plots has been made. All these variables are going to be the most important to differentiate the sex. Looking at the figure 6.20 it can be seen that FPT, FPTSexPD and FPTSexHC share the two most important features (“AGE” and “SDMTOTAL”). Other important features are “SCAU_total” and “UPSIT_total”. The same happens in FPISex, FPISexPD and FPISexHC. Figure 6.20 shows that the most important features are always “AGE”, “SDMTOTAL” and “MCAVFNUM”.

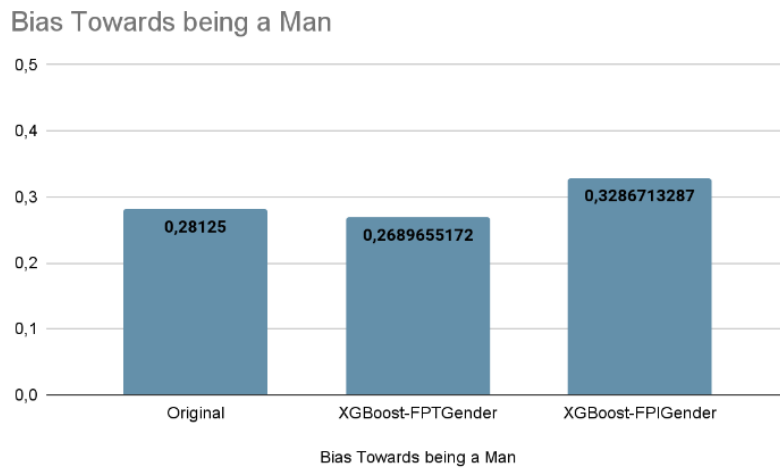
Although "AGE" is one of the most important attributes for predicting whether a subject has PD or is a HC subject, the fact that "AGE" is also very important in classifying the sex



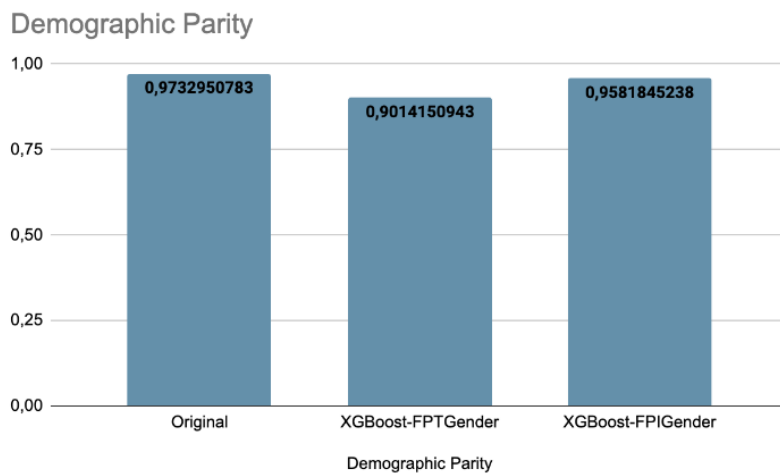
6.17 Figure: Confusion matrices for HC.

does not make much sense.

In this case the sex is not very well predicted. That is why, both in subjects with PD and HC subjects, with the tests carried out in this work, it has not been possible to find attributes to be able to make a correct classification.

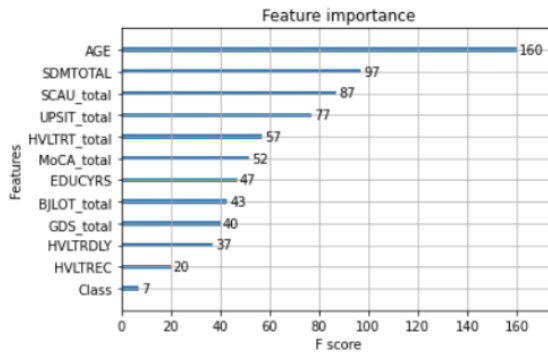


6.18 Figure: Demographic parity for sex classification.

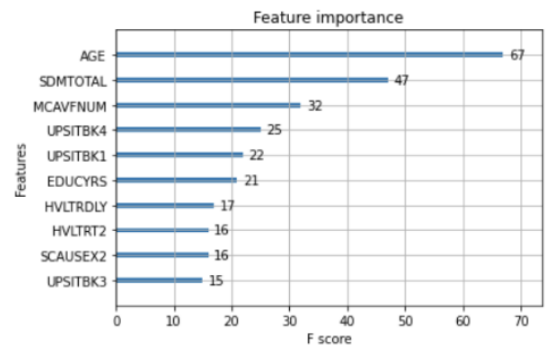


6.19 Figure: Bias towards being a man.

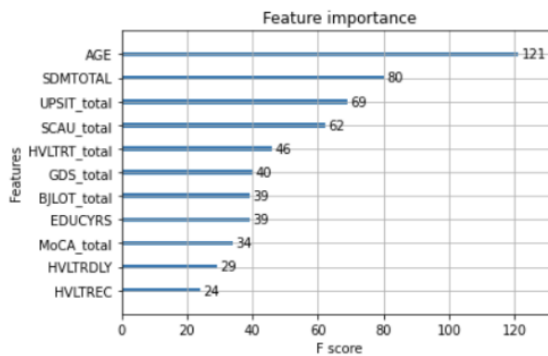
FPTSex



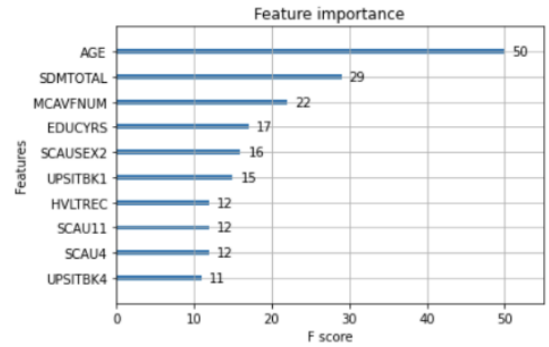
FPISex



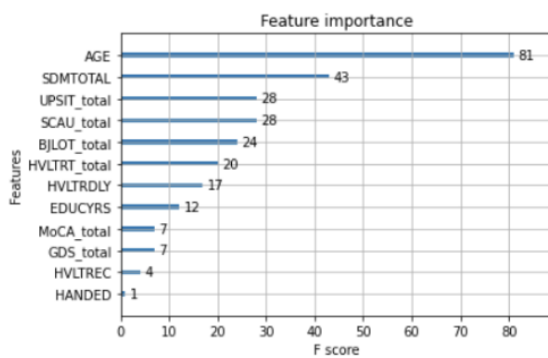
FPTSexPD



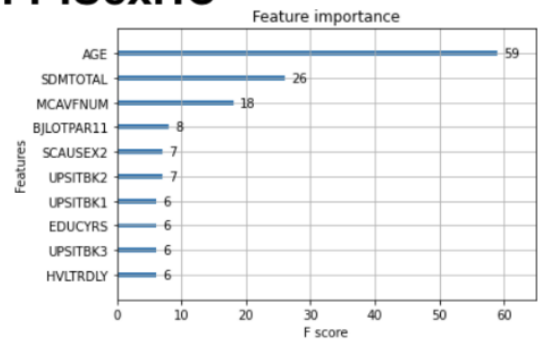
FPISexPD



FPTSexHC



FPISexHC



6.20 Figure: XGBoost ranking for sex classification.

7. CHAPTER

Conclusions

This project has analysed the influence of sex in diagnostic classification of Parkinson's disease based on non-motor symptoms by using machine learning methods. The impact has been evaluated using different metrics. The main analysis has been carried out using PPMI database. The data has been organised in different ways to test these machine learning methods in different scenarios in order to reach a better conclusion of the results obtained.

After having analysed the results obtained in different databases, taking into account the PD/HC classification, sex-specific classifiers seem to perform slightly better than global classifiers when it comes to classification. On the other hand, PD examples are better classified than HC examples.

Analysing the importance of the features, both SHAP and XGBoost tools agree on the selected variables. However, sex does not seem to be a determining factor in classifying between PD and HC.

Referring to the classification of sex, it seems that sex cannot be reliably classified according to the data obtained and the tests carried out. This conclusion has much to do with the fact that sex does not appear to be an important attribute in classifying PD/HC. That is why, both in subjects with PD and HC subjects, with the tests carried out in this work, it has not been possible to find attributes to be able to make a correct classification.

As a general conclusion, with these data and the methods used, for early Parkinson's patients the non-motor symptoms do not change according to sex.

7.1 Personal conclusions

Since I was very young, the world of healthcare has been very present in me, as both my mother and father work in this sector. I have always thought that IT can help to improve people's quality of life. That is why, thanks to this project, I have been able to witness in person how the union of the health world and IT make possible the development of a society where the quality of life of people increases by leaps and bounds.

It is true that there is still a lot of progress to be made and a lot of research to be done in these areas. However, the foundations have already been laid and that is why great advances will come very soon.

7.2 Future work

Once the project has been carried out, in order to improve it in the future, it would be useful to be able to perform performance analysis using other metrics such as "Precision", "Recall" and "F-measure". Although it is true that the graphs referring to these metrics have been extracted, they have not been exhaustively analysed due to the scope of the project. That is why they are included in the appendix so that they can serve as a reference in the future.

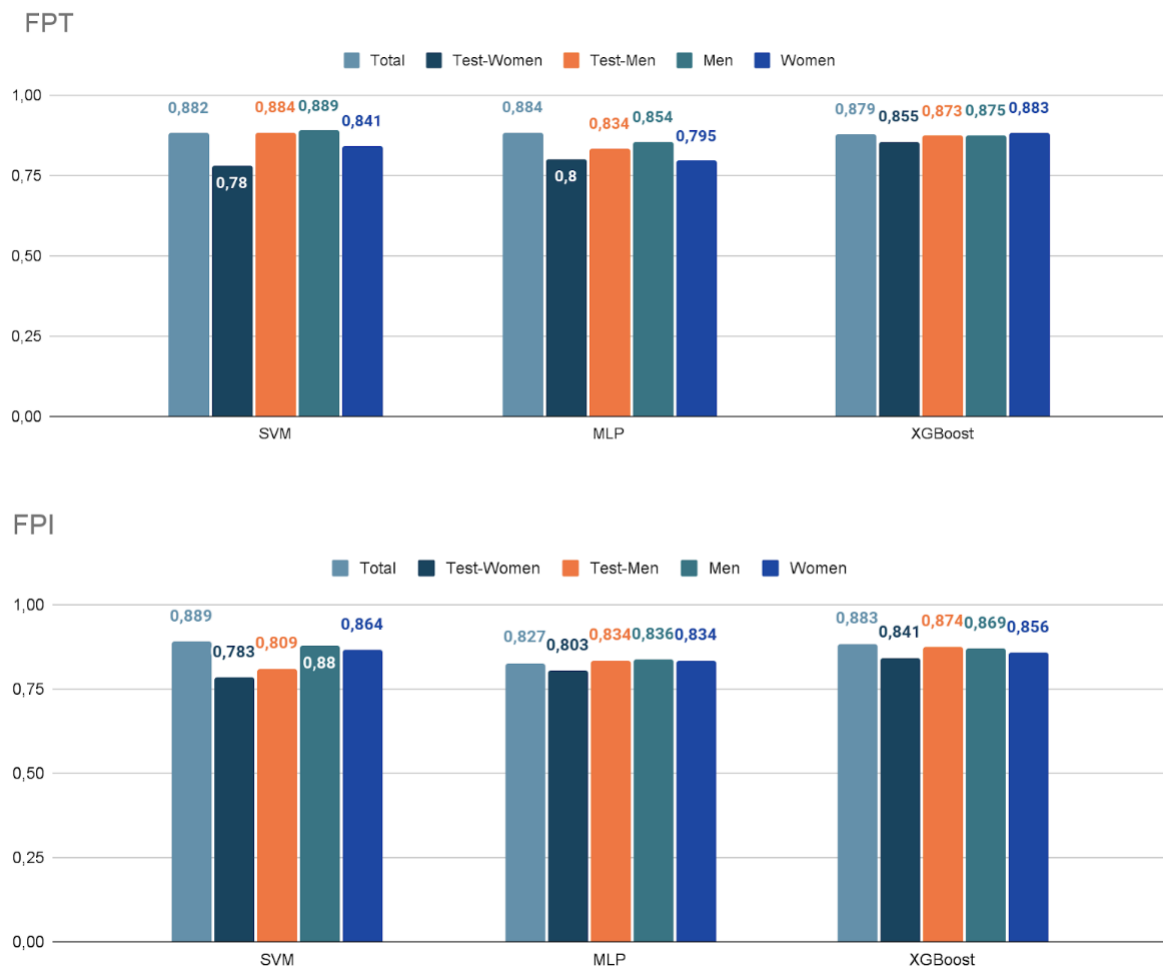
On the other hand, it would be interesting to have balancing samples. In the data obtained for the project it can be seen that the number of male subjects is much higher than that of female subjects and the same is true for PD and HC subjects. The difference is very noticeable and this is a reason to influence the classifiers. It would therefore be interesting to have synthetically-generated healthcare data [6] to solve this problem by preserving privacy and enabling researchers and policymakers to drive decisions and methods based on realistic data. And thus being able to test classifiers with balancing samples.

Appendices

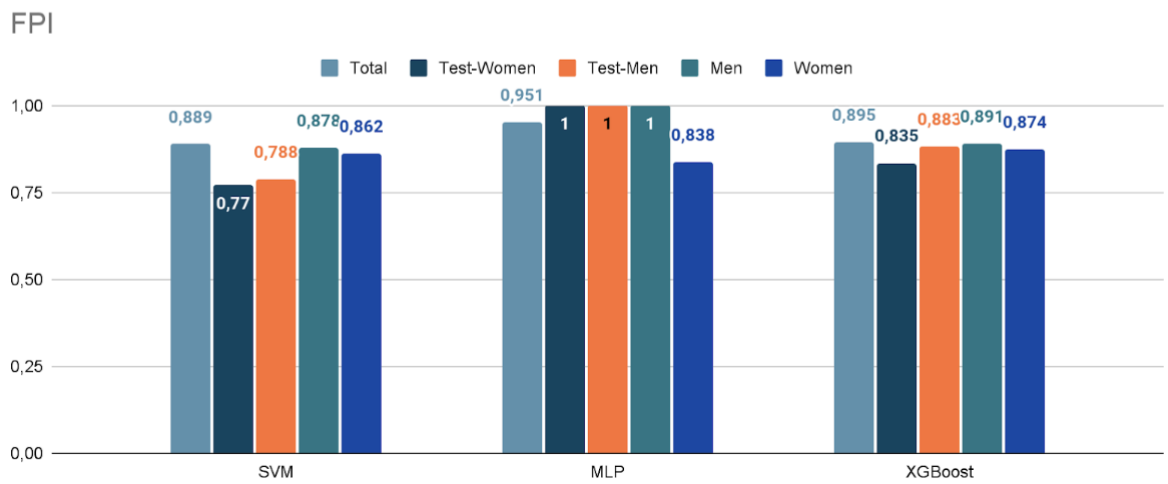
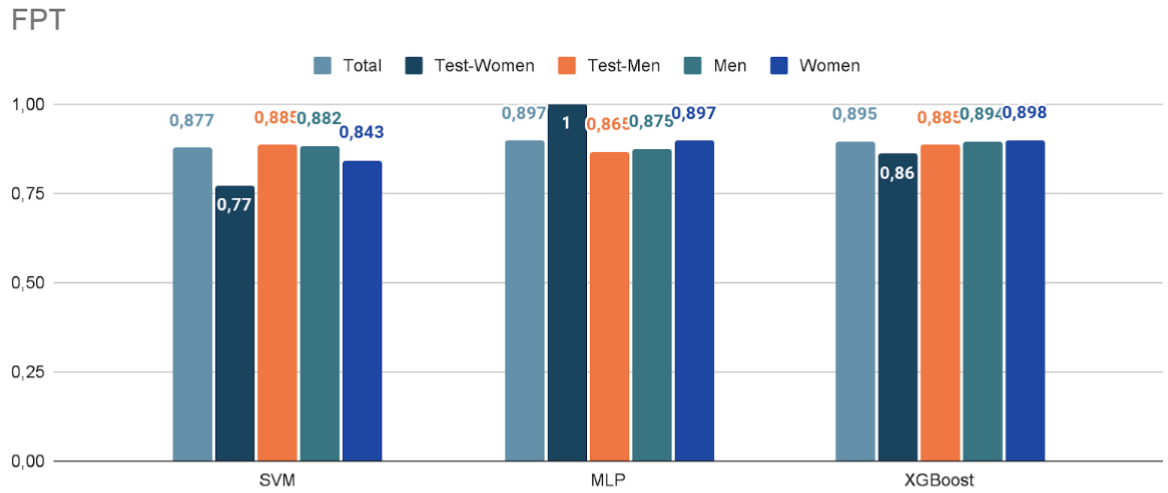
A. CHAPTER

Other metrics figures

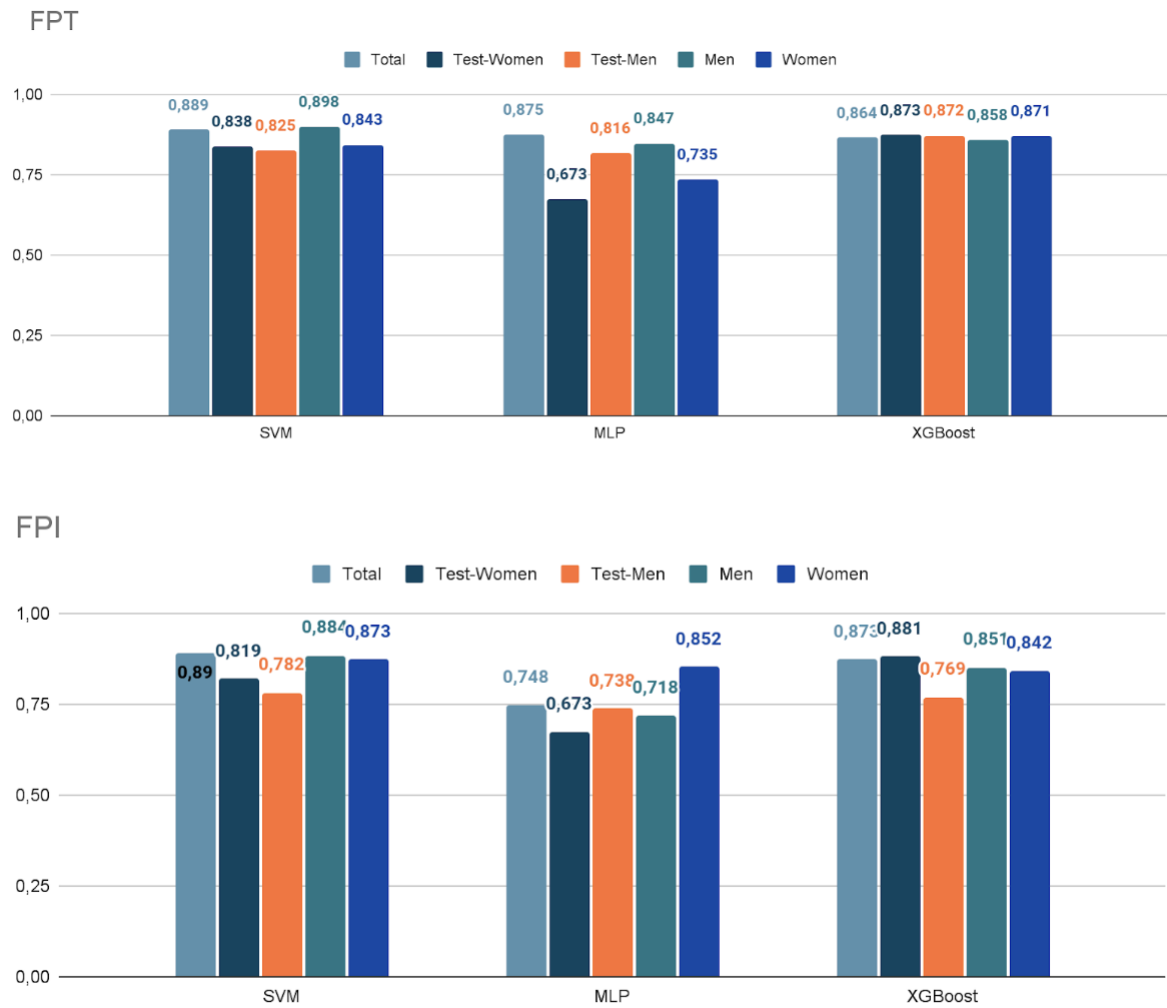
The graphs of different metrics obtained when classifying PD/HC and sex will be shown below. These metrics are F-measure, Recall and Precision. They are shown in this section because an exhaustive analysis has not been carried out and therefore they have not been taken into account in the conclusions. However, they may be useful for future projects related to this one.



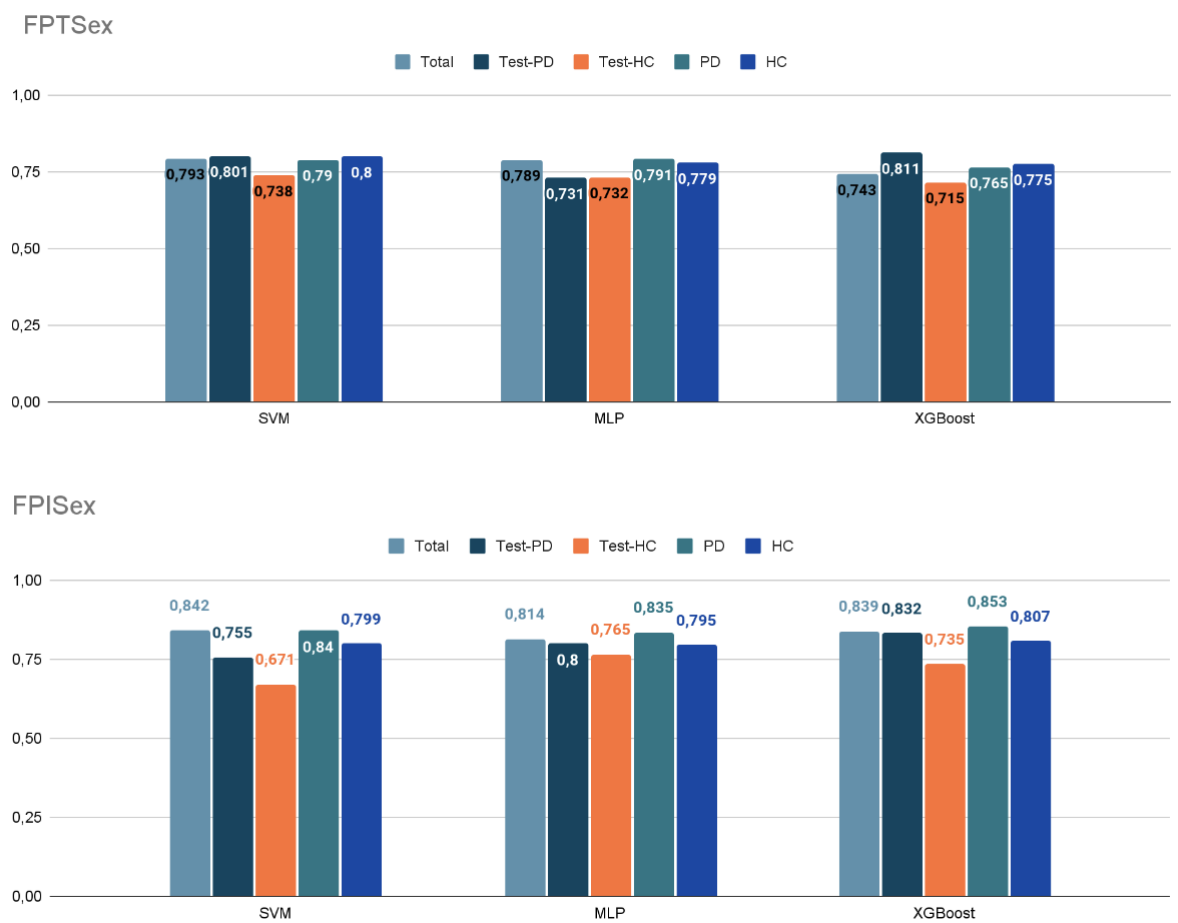
A.1 Figure: F-measure for FPT and FPI.



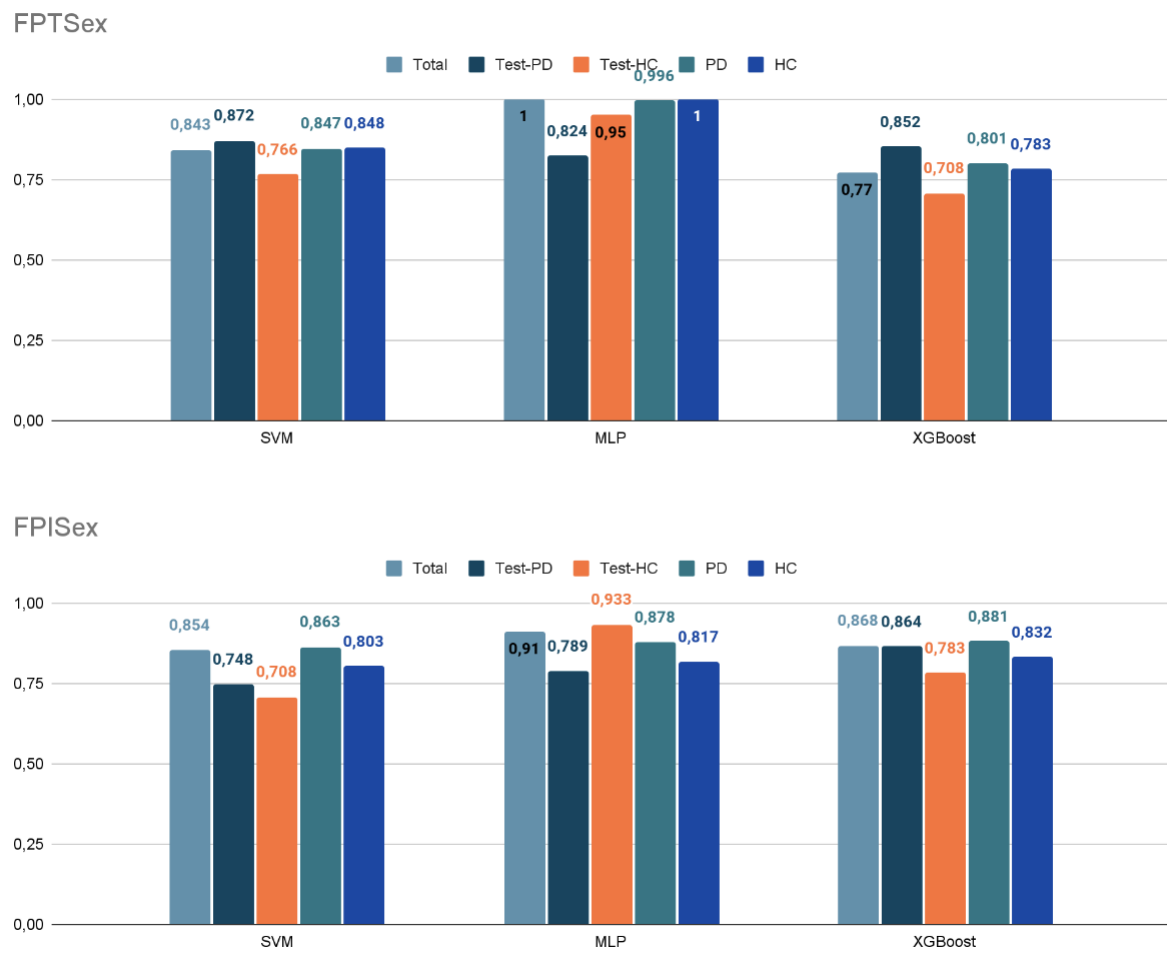
A.2 Figure: Recall for FPT and FPI.



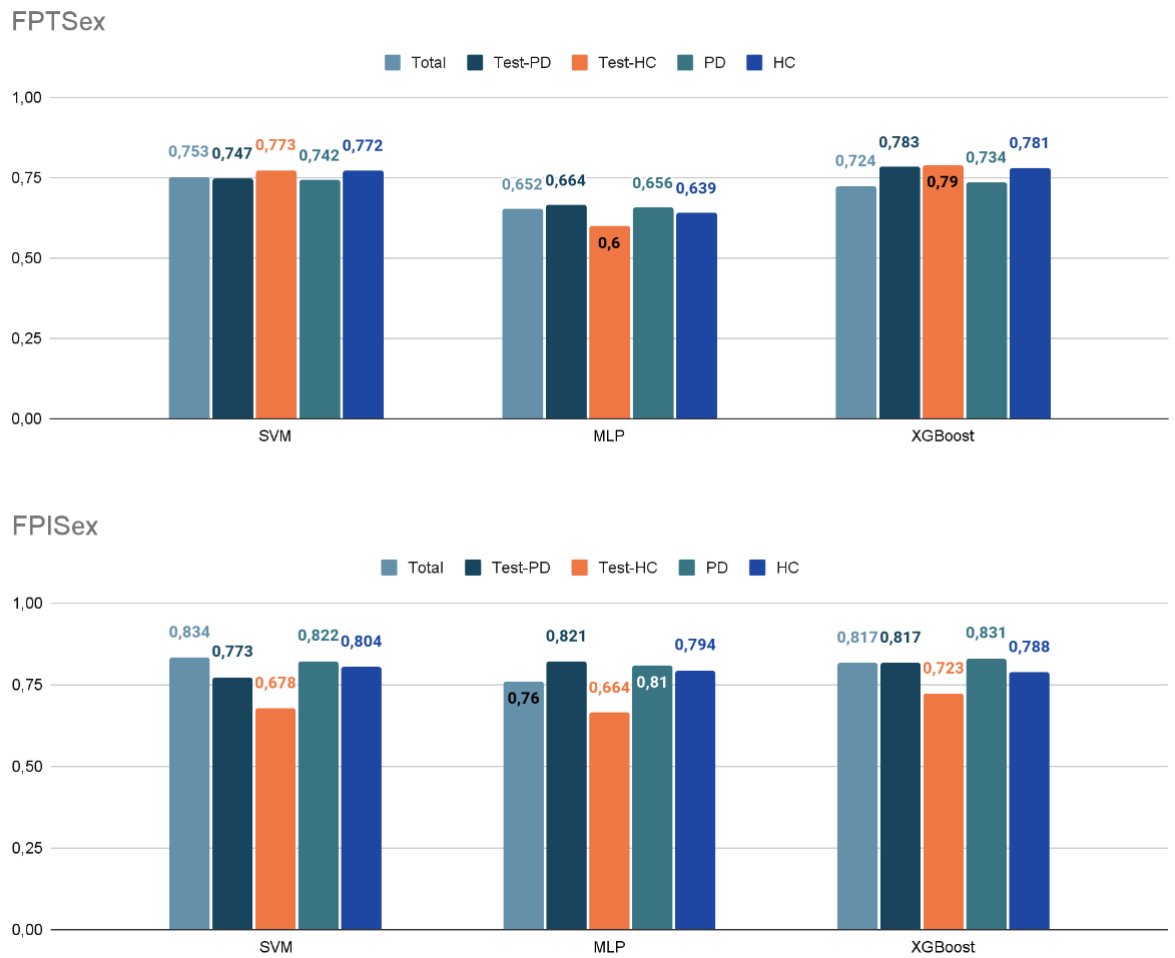
A.3 Figure: Precision for FPT and FPI.



A.4 Figure: F-measure for FPTSex and FPISex.



A.5 Figure: Recall for FPTSex and FPISex.



A.6 Figure: Precision for FPTSex and FPISex.

Bibliografia

- [1] Ahmad, A. S., Hassan, M. Y., Abdullah, M. P., Rahman, H. A., Hussin, F., Abdullah, H., and Saidur, R. (2014). A review on applications of ann and svm for building electrical energy consumption forecasting. *Renewable and Sustainable Energy Reviews*, 33:102–109.
- [2] Balestrino, R. and Schapira, A. (2020). Parkinson disease. *European Journal of Neurology*, 27(1):27–42.
- [3] Benedict, R. H., Schretlen, D., Groninger, L., and Brandt, J. (1998). Hopkins verbal learning test–revised: Normative data and analysis of inter-form and test-retest reliability. *The Clinical Neuropsychologist*, 12(1):43–55.
- [4] Bentéjac, C., Csörgő, A., and Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3):1937–1967.
- [5] Benton, A. L., Varney, N. R., and Hamsher, K. d. (1978). Visuospatial judgment: A clinical test. *Archives of neurology*, 35(6):364–367.
- [6] Bhanot, K., Qi, M., Erickson, J. S., Guyon, I., and Bennett, K. P. (2021). The problem of fairness in synthetic healthcare data. *Entropy*, 23(9):1165.
- [7] Bishop, C. M. et al. (1995). *Neural networks for pattern recognition*. Oxford university press.
- [8] Blanchard, G., Lugosi, G., and Vayatis, N. (2003). On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research*, 4(Oct):861–894.
- [9] Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168.

- [10] Char, D. S., Shah, N. H., and Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. *The New England journal of medicine*, 378(11):981.
- [11] Chauhan, V. K., Dahiya, K., and Sharma, A. (2019). Problem formulations and solvers in linear svm: a review. *Artificial Intelligence Review*, 52(2):803–855.
- [12] Chen, T. (2014). Introduction to boosted trees. *University of Washington Computer Science*, 22(115):14–40.
- [13] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.
- [14] Cheramy, A., Leviel, V., and Glowinski, J. (1981). Dendritic release of dopamine in substantia nigra. *Nature*, 289:537–42.
- [15] Das, R. (2010). A comparison of multiple classification methods for diagnosis of parkinson disease. *Expert Systems with Applications*, 37(2):1568–1572.
- [16] Dhaliwal, S. S., Nahid, A.-A., and Abbas, R. (2018). Effective intrusion detection system using xgboost. *Information*, 9(7):149.
- [17] Dolling, O. R. and Varas, E. A. (2002). Artificial neural networks for streamflow prediction. *Journal of hydraulic research*, 40(5):547–554.
- [18] Dong, B., Cao, C., and Lee, S. E. (2005). Applying support vector machines to predict building energy consumption in tropical region. *Energy and Buildings*, 37(5):545–553.
- [19] Dorogush, A. V., Ershov, V., and Gulin, A. (2018). Catboost: gradient boosting with categorical features support.
- [20] Dorsey, E. a., Constantinescu, R., Thompson, J., Biglan, K., Holloway, R., Kieburtz, K., Marshall, F., Ravina, B., Schifitto, G., Siderowf, A., et al. (2007). Projected number of people with parkinson disease in the most populous nations, 2005 through 2030. *Neurology*, 68(5):384–386.
- [21] Doty, R. L., Shaman, P., Kimmelman, C. P., and Dann, M. S. (1984). University of pennsylvania smell identification test: a rapid quantitative olfactory function test for the clinic. *The Laryngoscope*, 94(2):176–178.

- [22] Elglaly, A. K. A. (2018). Parkinson's disease and its management.
- [23] Freund, Y., Schapire, R., and Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612.
- [24] Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 28:337–407.
- [25] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.
- [26] Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378.
- [27] Gao, Y. and Sun, S. (2010). An empirical evaluation of linear and nonlinear kernels for text classification using support vector machines. In *2010 seventh international conference on fuzzy systems and knowledge discovery*, volume 4, pages 1502–1505. IEEE.
- [28] Gardner, M. and Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14):2627–2636.
- [29] Garnica-Caparrós, M. and Memmert, D. (2021). Understanding gender differences in professional european football through machine learning interpretability and match actions data. *Scientific Reports*, 11(1):1–14.
- [30] Graupe, D. (2013). *Principles of artificial neural networks*, volume 7. World Scientific.
- [31] Gumus, M. and Kiran, M. S. (2017). Crude oil price forecasting using xgboost. In *2017 International Conference on Computer Science and Engineering (UBMK)*, pages 1100–1103.
- [32] Jellinger, K. A. (2015). Neuropathobiology of non-motor symptoms in parkinson disease. *Journal of Neural Transmission*, 122(10):1429–1440.
- [33] Karlik, B. and Olgac, A. V. (2011). Performance analysis of various activation functions in generalized mlp architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems*, 1(4):111–122.

- [34] Kawamoto, K. R., Davis, M. B., and Duvernoy, C. S. (2016). Acute coronary syndromes: differences in men and women. *Current atherosclerosis reports*, 18(12):1–10.
- [35] Kingsford, C. and Salzberg, S. L. (2008). What are decision trees? *Nature biotechnology*, 26(9):1011–1013.
- [36] Lippmann, R. (1987). An introduction to computing with neural nets. *IEEE Assp magazine*, 4(2):4–22.
- [37] Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.
- [38] Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., Coffey, C., Kieburtz, K., Flagg, E., Chowdhury, S., et al. (2011). The parkinson progression marker initiative (ppmi). *Progress in neurobiology*, 95(4):629–635.
- [39] Martinez-Eguiluz, M., Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Perona, I., Murueta-Goyena, A., Acera, M., Del Pino, R., Tijero, B., Gomez-Esteban, J. C., et al. (2022). Diagnostic classification of parkinson’s disease based on non-motor manifestations and machine learning strategies. *Neural Computing and Applications*, pages 1–15.
- [40] Meng, Y., Yang, N., Qian, Z., and Zhang, G. (2020). What makes an online review more helpful: an interpretation framework using xgboost and shap values. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(3):466–490.
- [41] Mitchell, R. and Frank, E. (2017). Accelerating the xgboost algorithm using gpu computing. *PeerJ Computer Science*, 3:e127.
- [42] Moret, B. M. (1982). Decision trees and diagrams. *ACM Computing Surveys (CSUR)*, 14(4):593–623.
- [43] Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., and Chertkow, H. (2005). The montreal cognitive assessment, moca: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699.
- [44] Natekin, A. and Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21.

- [45] Niklason, G., Rawls, E., Ma, S., Kummerfeld, E., Maxwell, A. M., Brucar, L. R., Drossel, G., and Zilverstand, A. (2021). Explainable machine learning analysis reveals gender differences in the phenotypic and neurobiological markers of cannabis use disorder. *bioRxiv*.
- [46] Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567.
- [47] Nussbaum, R. L. and Ellis, C. E. (2003). Alzheimer’s disease and parkinson’s disease. *New England Journal of Medicine*, 348(14):1356–1364. PMID: 12672864.
- [48] Pal, S. K. and Mitra, S. (1992). Multilayer perceptron, fuzzy sets, classification.
- [49] Pinn, V. W. (2003). Sex and Gender Factors in Medical Studies: Implications for Health and Clinical Practice. *JAMA*, 289(4):397–400.
- [50] Poewe, W. (2008). Non-motor symptoms in parkinson’s disease. *European Journal of Neurology*, 15(s1):14–20.
- [51] Purushotham, S. and Tripathy, B. (2011). Evaluation of classifier models using stratified tenfold cross validation techniques. In *International Conference on Computing and Communication Systems*, pages 680–690. Springer.
- [52] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.
- [53] Roe, B. P., Yang, H.-J., Zhu, J., Liu, Y., Stancu, I., and McGregor, G. (2005). Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 543(2-3):577–584.
- [54] Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.
- [55] Ruder, S. (2016). An overview of gradient descent optimization algorithms.
- [56] Savica, R., Grossardt, B. R., Bower, J. H., Ahlskog, J. E., and Rocca, W. A. (2013). Risk factors for parkinson’s disease may differ in men and women: an exploratory study. *Hormones and behavior*, 63(2):308–314.

- [57] Shailaja, K., Seetharamulu, B., and Jabbar, M. (2018). Machine learning in healthcare: A review. In *2018 Second international conference on electronics, communication and aerospace technology (ICECA)*, pages 910–914. IEEE.
- [58] Sharma, S., Sharma, S., and Athaiya, A. (2017). Activation functions in neural networks. *towards data science*, 6(12):310–316.
- [59] Singh, N., Pillay, V., and Choonara, Y. E. (2007). Advances in the treatment of parkinson’s disease. *Progress in neurobiology*, 81(1):29–44.
- [60] Slack, D., Friedler, S., and Givental, E. (2019a). Fairness warnings and fair-maml: Learning fairly with minimal data.
- [61] Slack, D., Friedler, S. A., and Givental, E. (2019b). Fairness warnings and fair-maml: Learning fairly with minimal data. *CoRR*, abs/1908.09092.
- [62] Smith, A. (1968). The symbol-digit modalities test: a neuropsychological test of learning and other cerebral disorders. *Learning Disorders*, pages 83–91.
- [63] Štrumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665.
- [64] Vapnik, V.Ñ. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:774–780.
- [65] Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7.
- [66] Vyas, K., Vyas, S., and Rajyaguru, N. (2020). *Machine Learning Methods for Managing Parkinson’s Disease*, pages 263–294. Springer International Publishing, Cham.
- [67] Wang, Y., Liu, S., Wang, Z., Fan, Y., Huang, J., Huang, L., Li, Z., Li, X., Jin, M., Yu, Q., and Zhou, F. (2021). A machine learning-based investigation of gender-specific prognosis of lung cancers. *Medicina*, 57(2).
- [68] Wiens, J., Saria, S., Sendak, M. P., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K. A., Kale, D. C., Saeed, M., Ossorio, P.Ñ., Thadaneys-Israni, S., and Goldenberg, A. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, pages 1–4.

- [69] Wojtuch, A., Jankowski, R., and Podlewska, S. (2021). How can shap values help to shape metabolic stability of chemical compounds? *Journal of Cheminformatics*, 13(1):1–20.
- [70] Wooten, G. F., Currie, L. J., Bovbjerg, V. E., Lee, J. K., and Patrie, J. (2004). Are men at greater risk for parkinson’s disease than women? *Journal of Neurology, Neurosurgery & Psychiatry*, 75(4):637–639.
- [71] Wu, Q., Burges, C. J., Svore, K. M., and Gao, J. (2010). Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3):254–270.
- [72] Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., Adey, M., and Leirer, V. O. (1982). Development and validation of a geriatric depression screening scale: a preliminary report. *Journal of psychiatric research*, 17(1):37–49.
- [73] Zhang, Y. and Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, 58:308–324.
- [74] Zhao, H.-x. and Magoulès, F. (2012). A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, 16(6):3586–3592.
- [75] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints.
- [76] Zhao, W., Joshi, T., Nair, V.Ñ., and Sudjianto, A. (2020). Shap values for explaining cnn-based text classification models. *arXiv preprint arXiv:2008.11825*.
- [77] Zou, J. and Schiebinger, L. (2021). Ensuring that biomedical ai benefits diverse populations. *EBioMedicine*, 67:103358.