eman ta zabal zazu

Universidad
del País Vasco

Euskal Herriko
Unibertsitatea

INFORMATIKA
FAKULTATEA
FACULTAD
DE INFORMÁTICA

Master’s Thesis

Computational engineering and intelligent systems university master

Early detection of Parkinson’s disease based on non-motor symptoms

Maitane Martinez Eguiluz

Advisors
Gurrutxaga Goikoetxea, Ibai
Murueta-Goyena Larrañaga, Ane

July of 2021

Abstract

Background and objectives

Parkinson's disease (PD) is the second most common neurodegenerative disorder in the world, but the accuracy of clinical diagnosis is still limited, mainly in early stages when the cardinal motor symptoms are not present. This work aims to contribute to the early diagnosis of PD based on non-motor data from 490 patients with idiopathic PD and 197 control subjects. In addition, the most relevant biomarkers will be identified and the gender bias will be measured.

Methods

A database from an international repository (PPMI) was used, from which non-motor variables were selected. Four versions of the database with increasing granularity were generated, to which the Correlation Feature Selection method was applied to identify the most relevant variables for differentiating PD patients from controls. Then, eight classifiers were trained with machine learning algorithms (K-Nearest Neighbors, Support Vector Machine, Decision Tree, Consolidated Tree Construction, Naive Bayes, Multi-Layer Perceptron and Random Forest) and were statistically compared.

Results

Through these algorithms, a set of variables was detected that allowed differentiating patients from controls, suggesting that early detection of PD could be performed using a reduced version of the tests and questionnaires. The most relevant variables are related to impairments in the olfactory system. The algorithms with the best metrics were Support Vector Machine, Multi-Layer Perceptron and Random Forest, although only the second one achieved a balance between gender preconditions. In addition, using the explanatory algorithms, a set of simple rules capable of differentiating the two classes is proposed.

Conclusion

In this project, we explored the efficiency of several algorithms to differentiate between PD patients and healthy control subjects, using only non-motor characteristics. Olfactory impairment has been identified as the most relevant biomarker for this task and we propose a simple set of rules. Multi-Layer Perceptron algorithm, besides achieving good metrics, was barely affected by gender bias.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Neurodegenerative diseases are chronic and progressive disorders that affect the central nervous system, resulting motor and cognitive impairment. Neurodegenerative diseases can be broadly classified according to their clinical presentation and are typically defined by specific protein accumulations in vulnerable brain areas or cells.

1.1 Parkinson's Disease

Parkinson's disease (PD) is the second most common neurodegenerative disease after Alzheimer's disease [1] and it is characterised by the loss of dopaminergic (dopamine-producing) neurons in the substantia nigra [2, 3, 4]. The substantia nigra is a structure of the basal ganglia located in the midbrain. The disease was first described in 1817 by James Parkinson [5], and its prevalence is increasing: from 1990 to 2015, the number of patients with PD has duplicated in the world [6]. It is a disease that affects 1% of people over 60 years [7] and 15% of patients are diagnosed before the age of 40.

According to data from various epidemiological studies, there are between 80,000 and 100,000 PD patients in Spain and 8,000 new cases are diagnosed every year, i.e. there are 20 new cases per 100,000 inhabitants. Therefore, in Gipuzkoa there are around 1,000-1,400 people with PD and 140 new cases are diagnosed every year.

The origin of PD is unknown [8]. According to studies, it may be caused by two types of factors: genetic and environmental triggers. Inherited cases constitute a small percentage of PD, and the E46K point autosomal-dominant mutations in SNCA gene is restricted to a family in Biscay. They present a clinical phenotype that is characterized by early-onset and rapidly progressive parkinsonism followed by dementia. This family supervised by Biocruces Bizkaia Health Research Institute. Intriguingly, some carriers of E46K-SNCA mutation are asymptomatic or mild symptomatic, which poses questions about the precises epigenetic and environmental factors interacting with genetic risk factors to clinically develop the disease. Some known environmental risk factors for PD are exposure to chemical substances throught life or head trauma. Scientists have therefore determined that the reason for this disease is due to the interaction of both genetic and environmental factors.

The pathological hallmark of PD is a loss of dopaminergic neurons in mesencephalic areas. Loss of dopamine causes motor symptoms in PD patients, such as an involuntary resting tremor, slow movements, balance problems and rigidity [4, 1]. These cardinal symptoms appear after 50 to 70% of dopaminergic neurons are degenerated (called motor phase [9]), making early diagnosis difficult [1]. The diagnosis of PD is clinical and the presence of motor symptoms is essential for it. PD patients also present non-motor manifestations such as mood and sleep disorders, loss of smell, speech problems and nervous system dysfunction [7, 1]. These symptoms are known to develop years before PD diagnosis is made (called the prodromal phase [9]). It has been claimed that loss of smell might be present up to 20 years before diagnosis. Therefore, some non-motor symptoms might be helpful in detecting PD in prodromal stages, when neuronal loss is thought to be less pronounced [4].

Currently, the only available treatments for PD are for symptom relief, as there is no cure for this disease. One of the most commonly used drugs for this purpose is Levodopa (L-dopa) [5], which is a precursor of dopamine, and improves the motor status of the patients, reducing the slowness of movement and rigidity. For more advanced patients, when motor symptoms cannot be successfully controlled with pharmacological drugs, another treatment option is deep brain stimulation (DBS) therapy, which involves the implantation of a device to electrically stimulate certain regions of the brain [10]. However, there is an urge to develop therapeutic treatments that halt or slow down the progression of neuronal loss, that is, drugs that have a modifying effect on the course of the disease. For the success of such treatments, an early diagnosis is important.

In this regard, there is no specific analysis to diagnose PD. Neurologits experts in movement disorders rely on medical history, symptom analysis, and physical and neurological examination. Symptoms and signs of PD are detected using questionnaires and tests specific to the disease or specific to the symptoms. Physical examination is performed by disease-specific tests. The internationally validated examination for PD is the Unified Parkinson's disease Rating Scale or the UPDRS. The UPDRS evaluates various aspects of PD including non-motor symptoms, activities of daily living, motor impairment, and drug-induced complications. It includes a motor evaluation that characterizes the extent and burden of disease. The neurological examination is complemented with neuroimaging. Usually, single photon emission computed tomography (SPECT) technique is used as a confirmatory imaging test for PD diagnosis. This technique uses a radiotracer to identify the presence of dopaminergic deficiency in presynaptic terminals of basal ganglia.

Types of Parkinsonism

Parkinsonisms are a group of different clinical entities that share common symptoms reminiscent of PD, like slowness of movement, rigidity or tremor. Although the presentation of clinical symptoms is similar across parkinsonisms, the underlying cause varies and each diagnostic entity presents some characteristics that help to discriminate between conditions. These clinical entities that share common motor symptoms with PD are usually called atypical parkinsonisms, like Multiple System Atrophy, Corticobasal degeneration or Dementia with Lewy bodies. On the other hand, the differential diagnosis of Parkinson's disease should also include the following:

- Parkinson's disease (PD). The most common type of PD is the idiopathic PD (IPD), which accounts for 60-75% of cases [11]. Idiopathic means that the cause is unknown [12]. On the other hand, there is genetic PD, which are patients that are known to have genetic mutations that are risk factors for developing PD. In both cases, the characteristic symptoms are tremor, rigidity and slowness of movement [12], as well as a range of cognitive disorders, such as frontal lobe deficits [13, 14] or profound dementia [15].

- Vascular parkinsonism (VP) are produced by one or more small strokes, rather than by gradual loss of nerve cells as seen in the more typical neurodegenerative PD [16, 17, 18, 19]. The diagnosis is suggested by predominant involvement of the legs (“lower-body parkinsonism") with gait and balance problems, lack of tremor, poor response to Levodopa (as opposed to PD), and brain scans showing multiple minute or more extensive strokes [20].

- Drug-induced parkinsonism (DIP) is the most common movement disorder induced by drugs that affect dopamine receptors and is often misdiagnosed as IPD, as the symptoms are similar. Causative agents can be any medication that interferes with dopamine transmission, and mainly include olanzapine, risperidone, and aripiprazole. In addition, it can affect daily activities and can persist for long periods of time even after having stopped taking the drug that caused it [21].

- SWEED (Scans without evidence of dopaminergic deficit) patients are those with a diagnosis of Parkinson’s but no evidence of dopaminergic deficits. These patients have symptoms such as asymmetric rest tremor and absence of nigrostriatal dopaminergic pathway dysfunction.

1.2 Symptoms and diagnosis

1.2.1 Motor symptoms

The primary motor symptoms, considered cardinal symptoms of the disease, include akinesia, bradykinesia, rigidity, tremor y gait disturbances [22]. Akinesia is the loss of the ability to move muscles voluntarily. On the other hand, bradykinesia is slowness of movement, i.e. movements become slower and slower and, over time, the muscles may “freeze”. Rigidity, as the name suggests, is the muscle stiffness or inflexible muscles of the arms or legs beyond what would result from normal aging or arthritis.

Regarding tremor, the manifestations are differentiated between resting, postural and kinetic. Resting tremor is the most common of the three, while the other two, i.e. kinetic tremor (occurring during voluntary movements) and postural tremor, are more common in essential tremor [23]. Nevertheless, it is difficult to differentiate PD from essential tremor [24]. The tremor mainly affects the hands and feet, although other parts of the body may also be involved like jaw tremor, but to a lesser extent. 70% of PD patients experience tremor during the disease [25]. This tremor in hands has a characteristic presentation, named “pill rolling tremor", like if the patient was trying to roll a pill or another small object between their thumb and index finger.

Patients with PD are usually classified in studies into three categories: tremor-dominant, akinesia-dominant (also called akineto-rigid) or mixed phenotype category [26, 27]. Therefore, this subclassification of PD patients differentiates endophenotypes of PD. It is known that the likelihood of disease development might differ among PD patients and, for example, those with akinetic-rigid subtype might be more prone to suffer severe motor impairment compared to tremor-dominant subtype, which barely progress over time.

Gait disturbance is a common symptom. This affects PD patients who have difficulty lifting their feet off the ground, experiencing a shuffling gait.

In addition to the above symptoms, there are also some secondary symptoms such as impaired handwriting, speech and precision grip [22]. A large proportion of PD patients tend to have mycography, i.e. a form of tiny handwriting [28]. In addition, speech and voice disorders develop at some point in the disease in most PD patients [29]. Speech production is correlated with other motor symptoms, such as akinesia [30]. Finally, the precision grip is the grip formed by the index finger and thumb. It is used to evaluate loss of function in several disorders, such as PD [31].

1.2.2 Non-motor symptoms

The prevalence of sleep disorder ranges from 66% to 98% [32] and some of the abnormalities appear in the prodromal phase of the disease [33, 34, 35]. Sleep disorders such as insomnia [36], sleep fragmentation [37] or excessive daytime sleepiness [38] are common non-motor manifestations of PD. In addition, some patients might also present with restless legs syndrome [39] or rapid eye movement (REM) sleep behaviour disorder (RBD) [40].

Anxiety and depression are another two non-motor manifestations that are commonly present in PD, with a prevalence rate of 20-40% [41] and 50% [42], respectively. These symptoms are more prevalent in PD patients with REM sleep behaviour disorder [40], as they tend to score worse on anxiety and depression scales compared to healthy control subjects or even other PD patients [43]. These symptoms precede motor symptoms, as well as sleep disturbances [44].

Olfactory dysfunction also precedes motor symptoms and affects more than 90% of patients [45]. Some magnetic resonance imaging studies ([46, 47, 48]) have found that olfactory bulb volumes are significantly smaller in PD patients than in healthy control subjects. Myelin and axonal damage of the olfactory tracts has also been detected [49, 50], as well as pathological changes in the basolateral nucleus of the amygdala or in the anterior olfactory nucleus [51, 52]. Alterations in olfactory perception may also be due to changes in the connectivity of olfactory-related neural networks [53, 54, 55].

Visual functions are often affected in Parkinson's disease, such as decreased contrast sensitivity [56, 57]. In addition, colour discrimination is also impaired from early stages, but this does not occur in all PD patients, so it may represent a PD phenotype [58]. Color vision impairment has also been linked to dementia and cognitive impairment [59]. Moreover, some PD patients also suffer from visuoconstructive and visuoperceptual disturbances [60, 61, 62, 57, 63, 64, 65].

Regarding cognitive manifestations, PD patients are estimated to have a higher risk of developing dementia than healthy controls [66]. The manifestations precede the motor symptoms, with a prevalence of 20-25% for cognitive impairment level and 30% for dementia [7]. Symptoms include impaired decision-making, delayed verbal memory, slower processing speed and poorer attention [63, 67, 68, 69].

Finally, in addition to those mentioned above, gastrointestinal disorders are also manifestations of PD patients. The following symptoms are mentioned in [7]: hypersalivation

[70], dysphagia [71], nausea [72], gastroparesis [73], small intestinal dysfunction [74], slow transit constipation [75] and defecatory dysfunction [76]. As the other non-motor symptoms, gastrointestinal symptoms precede the onset of motor symptoms.

1.2.3 PD diagnosis

There is no specific test to diagnose PD. The physician trained in nervous system disorders (neurologist) will diagnose Parkinson's disease based on a medical history, a review of signs and symptoms, and a physical and neurological examination.

Neurological imaging examination

A nuclear medicine technique for imaging the dopamine transporter (DAT) is SPECT. This technique is often used to aid clinicians to confirm the diagnosis of PD [77]. Another technique is positron emission tomography (PET), although this technique is expensive and therefore may not be popular in clinical diagnosis [77]. However, both of these methodologies are based on detecting losses of dopaminergic neurons, so they are used to verify a diagnosis in the motor phase, but also to monitor the progression of the disease. It should be noted that patients can lose up to 80% of dopamine before symptoms appear, so these techniques might not be ideal for early detection of the disease [9]. On the other hand, magnetic resonance imaging (MRI) has been used to analyse structural changes in the brain and for the differential diagnosis of PD syndromes [77, 9]. Different MRI modalities are gaining attention in the last years for PD diagnosis, like neuromelanin-sensitive sequence or nigrosome 1 imaging, but these neuroimaging techniques have currently scarce clinical application.

Non-motor symptom questionnaires and tests

There are several types of questionnaires and tests that focus on detecting one or more non-motor symptoms. A questionnaire is a document in which a set of written questions are formulated and it is generally answered by choosing one of the options offered. The subject has to answer all the questions, sometimes assisted by the caregiver. A test refers to exercises designed to assess knowledge, skills or functions. The following are some of the questionnaires and tests that are often used to assess different aspects of PD:

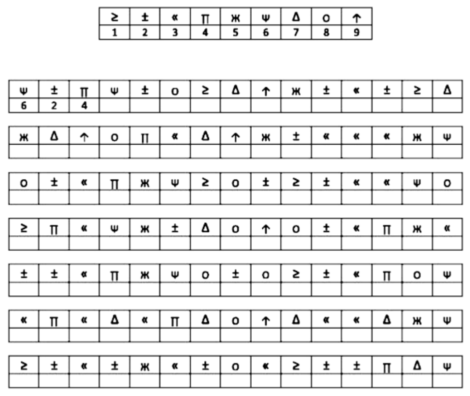
- Epworth Sleepiness Scale (ESS) [78] is a questionnaire with 8 questions, in which the subject qualifies the habitual possibilities of falling asleep while performing eight daily activities. It is measured on a 4-point scale, where 0 is never and 3 is a high probability. An example of one of the activities is watching TV or talking to someone.

- Geriatric Depression Scale (GDS) [79] represents a screening scale for depression on 15 yes or no questions, such as "Do you feel that your life is empty?". If the score is greater than 5, it suggests depression. The way of scoring is determined by the question, in other words, sometimes the answer "yes" indicates a symptom of depression and other times the answer "no".

- SCOPA-AUT [80] is a questionnaire about problems that have occurred in various bodily functions during the last month related to the autonomic nervous system; specifically, it consists of the following domains: gastrointestinal, urinary, cardiovascular, thermoregulatory, pupillomotor and sexual. 25 questions with 4 options are asked. The options are: “never",“sometimes", “regularly", “often" and in some questions there is an option of “not applicable".

- University of Pennsylvania Smell Identification Test (UPSIT) [81] is a test to evaluate the functioning of a person’s olfactory system. The evaluation consists of 4 different 10-page booklets, therefore it has a total of 40 questions. On each page, there is a strip of different microencapsulated scents. The subject has to scratch, sniff and identify the smell, choosing between 4 options. The results are the number of correct answers from each booklet.

- Symbol Digit Modalities Test (SDMT) [82] is a test for assessing neurological functions. This test consists of matching, in a limited time, 120 symbols with their corresponding numbers from 1 to 9, given as a reference a table that associates each digit with a different symbol, as shown in Figure 1.1. The result is the total of correct answers made before the time runs out.



≥	±	«	Π	ж	Ψ	Δ	o	↑
1	2	3	4	5	6	7	8	9

Ψ	±	Π	Ψ	±	o	≥	Δ	↑	ж	±	«	±	≥	Δ
6	2	4												

ж	Δ	↑	o	Π	«	Δ	↑	ж	±	«	«	«	ж	Ψ

o	±	«	Π	ж	Ψ	≥	o	±	≥	±	«	«	Ψ	o

≥	Π	«	Ψ	ж	±	Δ	o	↑	o	±	«	Π	ж	«

±	±	«	Π	ж	Ψ	o	±	o	≥	±	«	Π	o	Ψ

«	Π	«	Δ	«	Π	Δ	o	↑	Δ	«	«	Δ	ж	Ψ

≥	±	«	±	ж	«	±	o	«	≥	±	±	Π	Δ	Ψ

Figure 1.1: Example of SDMT.

- Benton Judgment of Line Orientation Test (BJLOT) [83] is a neuropsychological test that evaluates visuospatial judgment, through the task of discriminating the direction of lines. The test consists of two possible versions of 30 items. In each item there is a set of 11 lines with different angles that draw a fan which is open 180º. In addition, two lines of this set appear separated from the fan. Two examples can be seen in Figure 1.2. The goal is to match the two independent lines with their respective pair on the fan. The answer is considered correct only if the two lines have been paired correctly.

- Montreal Cognitive Assessment (MoCA) [84] is a screening test used to detect cognitive impairment. It assesses several cognitive domains: short-term memory, spatio-temporal reasoning skills, executive functions, attention, concentration and working memory, language, abstraction, reasoning and orientation to time and place. In total, all the tests add up to a maximum of 30 points. If the subject has 12 years of education or less, an extra point is added. A score of 26 or more is considered not to have any cognitive impairment. For example, spatio-temporal reasoning abilities are assessed using a clock-drawing task, which adds up to a maximum of 3 points.

- Hopkins Verbal Learning Test - Revised (HVLTR) [85] is a evaluation of verbal learning and memory. Each test has four nouns for each category, for a total of three

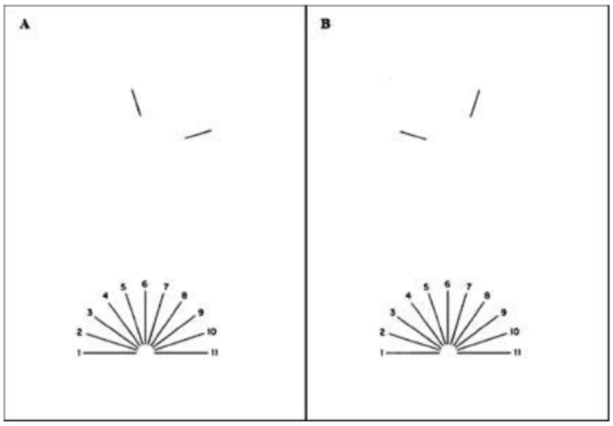


Figure 1.2: Example of BJLOT.

categories. Therefore, each test contains 12 words. An example of categories could be: temporary, instruments and bladed weapons. There will be three learning trials to learn the words, where the words learned in each trial will be verified. Approximately 20-25 minutes later, a deferred recovery test and a recognition test are completed. The first test requires free recall of any remembered words. The recognition test consists of 24 words, including 12 target words (from the initial list) and 12 false positives (6 semantically related and 6 semantically unrelated). In this test the identified words are recorded, but also the number of words that are not identified in the initial list, but are in the same category and also those that are semantically unrelated.

- Trail Making Test (TMT) [86] is a test used to assess attention, cognitive flexibility and visuospatial ability. It consists of two parts: in the first, the subject has to quickly join the numbers with lines, these being randomly placed in numerical order and in the second, the subject has to join the numbers and letters with lines, these being placed randomly, for example by joining the 1 with the A, the 2 with the B and so on. The result of both parts corresponds to the time (in seconds) that the subject has taken to complete the task. A short version of TMT is included in the MoCA test.

Motor symptoms questionnaires and test

A specific scale for PD is available for motor symptoms: Unified Parkinson's Disease Rating Scale (UPDRS). The UPDRS [87] is a combination of a questionnaire and a test that evaluates various aspects of PD. Among this aspects are the non-motor symptoms, activities of daily living, motor impairment and drug-induced complications. The questionnaire is divided into 4 parts:

1. Non-motor symptoms
2. Activities of daily living (related to motor impairment)
3. Motor examination
4. Levodopa-induced complications

The first two parts are questions related to everyday experiences, such as “Over the past week, have you had trouble with urine control?" or “Over the past week, do you usually have trouble turning over in bed?". In the third part, the examiner observes and evaluates some aspects of the patient, such as language problems or loss of facial expression. In addition, the examiner assesses motor activity through exercises, such as tapping the fingers as quickly as possible, walking 10 meters or extending the arms to observe the trembling of the hands. In the last part, three types of complications are evaluated: dyskinesias (unpredictable involuntary movements), motor fluctuations (changes in response to medication) and digestion related complications. This part only applies to patients with medicated PD.

The scale used for this questionnaire is the following: “Normal", “Slight", “Mild", “Moderate" and “Severe" and each level is scored as 0,1,2,3,4, respectively. Therefore, the higher the score, the higher disease severity.

1.3 Parkinson’s Progression Marker Initiative Data

The Parkinson’s Progression Markers Initiative (PPMI) is a study that aims to identify biomarkers for the progression of PD to improve therapeutic and etiological research. The identification of successful biomarkers will enable the improvement of therapeutic trials that could potentially modify the course of the disease [88]. The study is a public-private partnership funded by The Michael J. Fox Foundation for Parkinson’s Research (MJFF). The collected biomarkers are different in nature like imaging, genetic, biospecimen, or clinical data biomakers. Biomakers can be prognostic biomakers (those that can allow to track the progression of the disease) or diagnosis biomarkers (those that can help to differentiate PD patients from healthy subjects).

Two types of subjects from 24 study sites were selected for the current Master’s Thesis work: patients with PD and healthy control subjects (HC). Both types of subjects were of similar age and sex. At the time of enrollment, subjects with PD had to be at least 30 years old and had not received pharmacological treatment for the disease, among other requirements [89]. In addition, these subjects underwent dopamine transporter (DAT) imaging to check if there was a lack of dopamine (a characteristic symptom of the disease). Regarding healthy control subjects, at the time of enrollment they required an age of 30 years or more without an active neurological disorder.

All subjects gave written consent for clinical testing and neuroimaging prior to participation, approved by the Institutional Review Boards (IRB) of all participating institutions. After that, they underwent various tests: clinical, imaging evaluations, collection of biological samples from blood, urine, and cerebrospinal fluid. These evaluations were performed at the baseline (when the subjects were recruited) and every 3 months during the first year and every 6 months thereafter [88].

This database is free for researchers, with 400 subjects with recently diagnosed PD and 200 healthy subjects. Both types of subjects in PPMI will undergo a full longitudinal program of clinical assessments and imaging, as well as biospecimen collection, and non-motor testing.

1.4 Current data mining approaches for PD detection

The use of data mining techniques has increased in a number of areas over the last two decades [90]. In the medical field, specifically, the World Health Organisation (WHO) identified in 1997 the potential of using these techniques to improve medical diagnosis and prediction using medical data repositories [91]. More recently, psychiatry has also started to use these techniques in the area of mental disease, with the aim of better understanding the genetic composition [90].

Data mining is a process that works with a database in an attempt to discover patterns and knowledge. To do this, it uses machine learning algorithms and creates models that interpret the data in a useful way [92]. Machine learning algorithms are divided into 2 classes: supervised learning and unsupervised learning. Supervised learning infers prediction rules from data and applies them to new data, while unsupervised learning, groups data according to similarity and discovers patterns in the data [90]. There is a third class called semi-supervised learning which is a mixture of the two previous classes [93]. These algorithms have been used in the literature in the context of PD.

Papers using machine learning techniques are varied and can be classified by the type of task, the nature of the data, the techniques or the databases. For example, if we focus on the type of task, some of the projects focus on classifying between PD patients and HC subjects, while others try to find subtypes within PD or try to detect stages of the disease. In this section we divide the research according to the nature of the data, differentiating projects that use neural imaging-related data, articles that only use motor symptom features and publications that use data related to voice and speech signals.

Regarding the type of techniques, some of the algorithms are repeated independently of the data they use, such as Support Vector Machines (SVM), but other algorithms are data-specific, for example Convolutional Neural Network (CNN) for images. As for the preprocessing of neural images, each paper presents a different proposal, as well as the preprocessing of voice recordings. In addition, some of the papers present feature selection techniques together with the machine learning algorithms.

Finally, the articles mentioned below use different databases, although the most common one is PPMI[1]. On the other hand, many of the articles in the voice and speech signals section use Parkinson's Data Set[2], which contains several voice signals.

Neuroimaging-based classifications

Many of the works focus on discriminating PD patients from HC subjects by performing an imaging-based neurological examination. For this purpose, some researchers use MRI imaging ([77, 6]), while others use DAT SPECT imaging ([3, 94, 95]). As far MRI imaging is concerned, Adeli et al. [77] use a joint feature-sample selection (JFSS) together with a Robust Linear Discriminant Analysis (RLDA) classifier, although Yasaka et al. [6] propose to apply a CNN classifier to different areas under the receiver operating characteristics curve. As for the DAT SPECT imaging, Oliveira et al. [3] used three classifiers: SVM, K-Nearest Neighbors (KNN) and Logistic Regression (LR). The authors reported that with SVM they have had better results. Wenzel et al. [94] propose to use a CNN, but Llera et al. [95] apply a probabilistic normalisation based on a mixture of Gamma distributions.

[1] https://www.ppmi-info.org/

[2] https://archive.ics.uci.edu/ml/datasets/parkinsons

Other studies combine imaging data with clinical data. For example, Singh et al. [96] use MRI images together with the UPDRS, MoCA, BJLOT and GDS tests to differentiate between PD patients, HC subjects and SWEED patients. To do so, they use Principal Component Analysis (PCA) for feature extraction, Fisher Discriminant Ratio (FDR) for feature selection and SVM for classification. In Amoroso et al. [9], MRI images are combined with clinical features (Age, ESS, GDS, MDS-UPDRS, MoCA and RBD tests) to make the binary classification between PD and HC. For this purpose, they use Random Forest (RF) as a feature selector and SVM as a classifier.

In addition to the above, Rahmim et al. [2] combine MRI and DAT SPECT imaging features with other clinical features to predict motor severity in PD patients in four years. They use the MDS-UPDRS - part III scale in year 4 and the clinical features are demographics, disease duration, UPDRS-III motor measures and MoCA. To analyse the images, they performed automatic region-of-interest (ROI) extraction on the MRI images, registered the SPECT images onto the corresponding MRI images and extracted the radiomic features. Once they had all the features, they applied RF. Castillo-Barnes et al. [8] use DAT SPECT imaging with biospecimen analysis results to differentiate HC subjects from PD or SWEED patients. For this purpose, they present an Ensemble Classification model with Performance Weighting, in which they use several SVM classifiers with linear kernel in different groups of biomedical tests, such as Cerebrospinal Fluid, RNA or Serum.

In addition to using neural imaging and other features to differentiate PD patients from HC subjects, other approaches have also been tried. On the one hand, Si-Chun et al. [97] attempts to predict depression in PD patients by combining demographic parameters, clinical parameters, cerebrospinal fluid levels and DAT SPECT images. To do so, they used the extreme gradient boosting (XGBoost) algorithm and logistic regression technique to predict the scale of GDS.

On the other hand, motor and non-motor tests, biospecimen examinations and neuroimaging results have also been combined to detect Parkinson's subtypes. Zhang et al. [98] use this data collected over 6 years and, using Long Short-Term Memory (LSTM), create a multidimensional time series for each patient with Idiopathic PD. This results in 3 subtypes of PD: the first subtype is characterised by stable cognitive ability but moderate motor impairment; the second subtype is characterised by moderate motor and non-motor impairment; the third subtype is characterised by rapid progression of motor and non-motor symptoms.

Movement-based classifications

As mentioned in the introduction, PD patients suffer from motor symptoms and these symptoms are also used in Parkinson related machine learning tasks. One of the most common ways of measuring the severity of these symptoms is using the third part of the MDS-UPDRS test. For example, Cavallo et al. [99] acquire data on upper limb movement by performing six MDS-UPDRS III tasks. Then, they apply and compare three classifiers (SVM, RF and Naive Bayes (NB)) on different datasets. The best result is obtained with the RF classifier. There are also other ways of measuring movement, such as gait analysis as proposed by Abdulhay et al [100]. In this article they extract gait characteristics to which a SVM classification is applied. The characteristics are the following: stride time, stance time, swing time and foot strike profile.

Other researchers propose to differentiate PD patients from HC subjects by drawing movement tests. In Kotsavasiloglou et al. [101] article, subjects are asked to draw horizontal lines on a tablet. In this test, characteristics such as velocity variability, deviation from the horizontal plane and trajectory entropy are collected. Once the features are obtained, they combine various feature selection methods with different machine learning techniques. Gupta et al. [102] proposed a different way: to follow the line of a spiral and a meander. With such handwritten tests, the difference between the tracing and the template can be obtained. Then, they apply an optimised crow search algorithm combined with KNN, RF and DT.

On the other hand, Kuhner et al. [103] try to identify movement characteristics that differentiate HC subjects from PD patients. To do this, they use deep brain stimulation (DBS) of the subthalamic nucleus (STN) turned off and on. A 10-metre walk is evaluated by monitoring values of position, velocity, acceleration and jerk vectors of segments and joints. To find the most discriminating features they use AdaBoost.

Voice and speech signals

Another common way to perform the binary classification task is by using voice signals. The voice signals are usually vowel phonations ([4, 104, 105, 106, 107, 108, 109]), although in some works other recordings are also used, such as speech with the pronunciation of a short phrase in Lithuanian language ([104]). From these signals, characteristics related to frequency, amplitude and pitch, i.e. acoustic characteristics, are extracted. The best performing classifiers are SVM ([4, 107]), KNN ([104, 105]) and Neural Networks ([110]). In Wang et al. [106], a method for classification called AABC-KWELM is proposed, which deals with class imbalance. Naranjo et al. [108] developed a two-stage Bayesian selection and classification approach. Finally, Wodzinski et al. [109] calculate the spectrum of the audio recordings to convert them to images and use them as input for the pre-trained ResNet architecture.

Another approach to the use of speech signals is given by Nilashi et al. [93]: they try to relate the properties of the speech signal and the UPDRS scores. Some of the properties are measures of frequency variation, measures of amplitude variation or measures of the relationship between the noise and the tonal components of the voice. In this article four different classification schemes are used and compared: Neural Networks, DMneural, Regression and Decision Tree.

Non-motor symptoms for PD detection

No studies or articles related to machine learning using only non-motor features have been found. In some of the articles mentioned above these data have been used together with other data, such as MRI images [96]. Therefore, given that non-motor symptoms are present in early stages of PD, there is an urgent need to explore the potential usefulness of non-motor symptoms in early PD diagnosis.

CHAPTER 2

Objectives

Parkinson's disease (PD) is the second most common neurodegenerative disorder in the world. Characteristic symptoms of PD are motor symptoms, such as involuntary resting tremor, slow movements or balance problems appear after neurodegeneration of mesencephalic neurons has undergone for years. However, PD patients develop non-motor manifestations during disease development, even preceding the onset of motor symptoms. Therefore, non-motor symptoms could be useful for early PD diagnosis. This work has the hypothesis that machine learning models for supervised classification, created from non-motor features, are able to efficiently discriminate healthy control (HC) subjects from patients with PD.

In order to test this hypothesis, medical and data mining knowledge is required. This knowledge is provided by the two research groups that support this project and co-direct the master's thesis. Both parts have defined a series of objectives:

1. Create machine learning classifiers to differentiate between HC and PD.
2. Identify the most relevant biomarkers.
3. Examine the gender effect.

To achieve these objectives, the following tasks have been defined to be fulfilled:

- **Technology watch.** Make an exhaustive and continuous bibliographic review and follow-up of the main conferences in the area.
- **Data obtention and pre-process.** The main source of data for this project will be the PPMI database, where a lot of data related to PD patients and HC subjects can be found. This database will be analyzed to obtain a statistical description and different versions of the database will be created in order to analyze which level of abstraction is the most suitable to achieve the proposed objectives.
- **Detection of Parkinson's disease.** The objective is to build classifiers that efficiently discriminate between PD patients and HC subjects based on non-motor

symptoms and on the identification of the most relevant features in the process. First of all, feature selection algorithms will be used to identify the most relevant features to discriminate PD patients from HC subjects and, then, several classifiers will be trained with the available data and multiple quality metrics (accuracy, precision, recall...) will be measured. Finally, a statistical comparison of the models obtained will be performed.

- **Biomarker detection.** Using the results of feature selection algorithms, together with explanatory models such as decision trees or rule-based algorithms, the most relevant non-motor biomarkers for differentiating patients with PD will be detected.
- **Adapt the experiments to the gender perspective.** It is scientifically proven that gender influences on PD symptoms and signs, but very few of the studies take this into account, as they do not carry out any studies separating the data according to gender. Therefore, it will be analyzed whether the classifiers constructed are adequate to deal with each gender or whether separate databases and classifiers should be created.

CHAPTER 3

Methodology

In this chapter, the methodology used in the project is presented. First of all, the data used in the project is described. In addition, the preprocessing and statistical description performed are explained. Finally, the data mining process is explained.

3.1 Data obtention and pre-process

PPMI database is the main source of data for this project. In addition, the Biocruces Bizkaia Health Research Institute has also provided its own data collected in previous studies.

The data obtained from PPMI were analysed and pre-processed by Andoni Angulo Celada, in his master's thesis [111]. In the preprocessing, he created a database from the public PPMI data, by selecting the non-motor questionnaires and clinical test that coincided with the ones collected in the clinical studies performed in Biocruces Bizkaia HRI, so that both are compatible. Errors were also eliminated and the problem of missing values was solved.

A tabular database has been constructed using the previously preprocessed data. These data contain non-motor feature information along with the diagnosis (PD or HC) or class. This supervised database uses data on the first patient visit, which is why we have called it the baseline database. To build this database, several decisions have been made during the work, although only the final one is presented in this document.

Learning algorithms can be fed with individual data obtained from questionnaires and tests performed by experimental subjects, but this approach can lead to suboptimal results, because the data might be highly correlated and the feature space could be very high. On the other hand, by summarising each test and questionnaire to a single score, many distinctive attributes of the subjects can be hidden, rendering totals scores also suboptimal. In an attempt to examine which levels of information gathering would be ideal to differentiate PD patients from controls, four different versions of the database have been created, with the aim of finding if a middle ground between the two extremes.

Finally, the dataset has been analysed with statistical methods to get a clear picture of the nature and structure of the available data. In addition, this analysis will uncover poor quality data (outliers, redundant variables and so on) that need to be corrected or removed.

3.1.1 Baseline database

A total of 687 subjects were drawn from the PPMI database: 490 (71%) with idiopathic PD and 197 (29%) were HC. 34% were female and the remaining 66% were male. The attributes used in this work are divided into three groups: demographic and clinical, questionnaire results and clinical test results. The results obtained at each subject's initial visit were selected.

General and clinical attributes include: gender (GENDER), years of education (EDUCYRS), dominant hand (HANDED), age (AGE) and the class itself (PD or HC). The questionnaires and tests are as described in Section 1.2.3.

Table 1 in Appendix A summarises the variables of the different versions of the database. The first version of the database (called Individual version) contains the results of each item comprising the all questionnaire and clinical test questions, in addition to the general and clinical attributes mentioned above. The variables of Individual are grouped semantically in the intermediate versions, creating two different combinations of them. Finally, all intermediate variables from each test or questionnaire are grouped together to create a single variable. The operation used to group these individual items has been the sum. The number of features in each version can be seen in the Table 3.1.

Total	Intermediate 2	Intermediate 1	Individual
15	29	53	107

Table 3.1: Number of variables in each version of the supervised database.

3.1.2 Pre-processing

In this section, the pre-processing of the data will be explained. Data is often not clean, it contains missing values, outliers that add noise and correlation between variables. As mentioned above, missing values from both databases have already been identified and treated.

In addition to this, outliers have been identified and processed in this project. For example, in the questions related to sex in the SCOPA_AUT questionnaire, there is an alternative option “not applicable". All the other questions are scaled from 0 (“never") to 3 (“often"), so it has been considered more convenient for this alternative to take a value of 4. Therefore, it is a pseudo ordinal variable, because on the one hand it has a frequency scale and on the other hand the question is not applicable in a specific case. The same criteria was used to assign a value to the option “use catheter" in the same test. A detailed explanation of the values taken by each variable is given in Appendix A. In addition, the semantic groupings made are also explained.

Finally, when analysing the univariate descriptive variables, it was observed that they do not have the same range. Therefore, it has been decided to apply two preprocessing techniques when it has been required: One-hot encoding has been applied to categorical variables, while qualitative variables have been normalised by min-max normalisation. In this way, all variables take a value in the range [0,1].

One-hot encoding

One-hot encoding is the transformation of categorical variables into binary vectors. For each category, a new variable is created and indicates by 1 whether that sample belongs to that category.

For example, from the variable GENDER the variables GENDER0 (woman with reproductive capacity), GENDER1 (woman without reproductive capacity) and GENDER2 (man) are created. GENDER0 will take a value of 1 if GENDER takes a value of 0 and the other two variables will take a value of 0, but, if GENDER takes a value of 1 then the one that will take a value of 1 will be GENDER1.

Rescaling (min-max normalization)

Min-max normalization rescales the range of values of the features to scale them to the range [0,1]. For these, the following formula is applied:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{3.1}$$

where x is the original value, x' is the rescaled value.

3.1.3 Data analysis

In this section the methods used for the analysis of the variables are explained, including univariate and bivariate descriptive statistics. In addition, several graphs will be added to be able to observe the behaviour of the variables.

In general, there are two types of variables, depending on the nature of the observation space:

- Quantitative. The observation space is a magnitude which is expressed by numbers (which have an algebraic structure by which they can be operated).
- Qualitative. The observation space is not a magnitude, but a category (it cannot operate between them). Categories can be coded by numbers, which is the usual way, but numbers do not express magnitudes, they express codes.

3.1.3.1 Univariate descriptive statistics

Univariate descriptive statistics have been analysed for the predictor variables. To do this, different types of graphs have been added: on the one hand, a bar chart and pie chart have been used to visualise qualitative variable; on the other hand, a box plot and a density histogram have been used to visualise the quantitative variables.

In addition, the means and standard deviations of all variables in the total version, divided by classes, as well as their p-values for the difference between classes have been analysed. Welch's t-test is used to calculate the p-value.

Box plot outlier

The outliers, i.e. the observations that are numerically distant from the rest of the data, are represented by a dot, as shown in Figure 3.1.

Welch's t-test

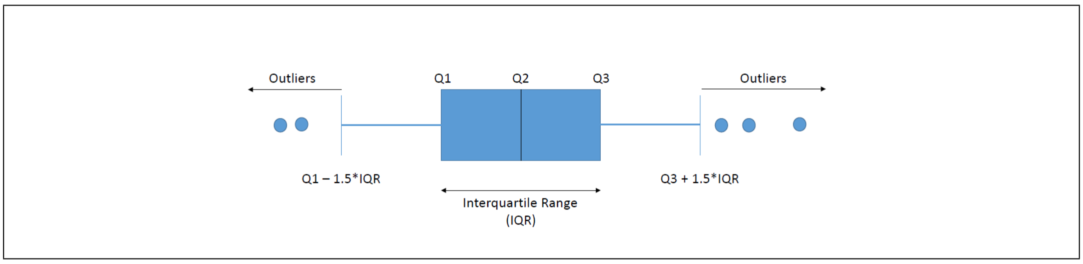


Figure 3.1: Box plot representation.

This test is an adaptation of the Student's test, in which it is assumed that the two samples do not have the same variance, but the assumption of normality is maintained. It is used to test the hypothesis of two populations having equal means. Mathematically, given two samples of size n_1 and n_2, one having mean $\bar{x}_1$ and standard deviation σ_1, the second having mean $\bar{x}_2$ and standard deviation σ_2, Welch's t-test defines the statistic t according to the following formula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{3.2}$$

3.1.3.2 Bivariate descriptive statistics

As for the bivariate descriptive statistics, we wanted to analyze the correlation of the variables between the different tests and questionnaires, as well as their correlation with the general and clinical attributes. For this reason, only the Intermediate 2 and Total versions have been analyzed, since they have fewer variables and the groupings provide a higher level of complexity. Furthermore, it has been taken for granted that test or questionnaire questions are correlated with each other, since these types of items are specific to detect certain symptom.

In order to calculate the correlation coefficient, the work of Harry Khamis [112] has been taken into account. In this article he differentiates variables into three scales of measurement: continuous, ordinal and nominal. Continuous variables express a numerical quantity, ordinal variables an ordered category and nominal variables an unordered category.

Looking at the variables in Intermediate 2 version, the nominal variables would be the following: GENDER, HANDED and Class. There is no ordinal variable, therefore continuous variables would be all other variables. For the total version, they would be the same variable classification. Taking into account the characteristics of the variables, the correlations indicated in Table 3.2 are applied.

In order to be able to apply the rank-biserial correlation coefficient and the point-biserial correlation coefficient, the nominal variables have been transformed into one-hot. In order to know the correlation between the new binary variables, the phi coefficient method will be applied to them. On the other hand, Goodman's and Kruskal's lambda correlation has been applied to the non-transformed nominal variables in order to see the real correlation between them. All these measures are explained below.

Pearson correlation coefficient

Pearson's correlation coefficient is a measure of linear dependence between two vari-

	Nominal	**Ordinal**	**Continuous**
Nominal	φ or λ	Rank biserial	Point biserial
Ordinal	Rank biserial	Kendall's τ_b	Kendall's τ_b
Continuous	Point biserial	Kendall's τ_b	Pearson

Table 3.2: Correlation coefficient to be applied for each type of variable. φ = phi coefficient. λ = Goodman and Kruskal's lambda.

ables. Moreover, this measure is independent of the scale of measurement of the variables. Mathematically, Pearson's correlation coefficient (rho) of two random variables (X and Y) is defined as

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}, \tag{3.3}$$

where Cov is the covariance of (X, Y) and Var the standard deviation of the variables. The coefficient returns a value in the interval $[-1, 1]$, where the sign indicates the direction of the relationship:

- $\rho_{X,Y} = 1$: there is a positive correlation in which the two variables have a direct relationship (if one increases, the other increases in constant proportion).
- $\rho_{X,Y} = 0$: there is no linear relationship.
- $\rho_{X,Y} = -1$: there is a negative correlation in which the two variables have an inverse relationship (if one increases, the other decreases in constant proportion).

Kendall's coefficient of rank correlation τ_b

Kendall's rank correlation coefficient, also known as Kendall's coefficient, is a measure of rank correlation. This coefficient is used when one or both measurement scales of the variables are ordinal. Therefore, it measures the ordinal association between two quantities. Mathematically, given two variables (X and Y), the Kendall τ_b coefficient is defined as

$$\tau_b = \frac{(P - Q)}{\sqrt{(P + Q + T) * (P + Q + U)}}, \tag{3.4}$$

where P is the number of concordant pairs, Q the number of discordant pairs, T the number of ties only in X, and U the number of ties only in Y. If a tie occurs for the same pair in both X and Y, it is not added to either T or U.

The interpretation of the τ_b values is the same as for the Pearson values, i.e. -1 means a negative association, 1 a positive association and 0 no association.

Point-biserial correlation coefficient

Point biserial correlation coefficient (r_{pb}) is a correlation coefficient used when one variable is dichotomous. The point biserial correlation is mathematically equivalent to the Pearson correlation, i.e. if we have a continuous measurement variable X and a dichotomous variable Y ($\rho_{X,Y} = r_{pb}$).

Mathematically, having a dichotomous variable Y (with values of 0 and 1) and a continuous variable X, the point biserial correlation coefficient is defined as

$$r_{pb} = \frac{M_1 - M_0}{s_n}\sqrt{\frac{n_1 n_0}{n^2}}, \tag{3.5}$$

where s_n is the standard deviation used when data are available for every member of the population:

$$s_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}, \tag{3.6}$$

M_1 is the mean value of the continuous variable X when Y is 1, and M_0 is the mean value of the continuous variable X when Y is 0. In addition, n_1 is the number of data points when Y is 1, n_0 is the number of data points when Y is 0 and n is the total sample size.

Rank-Biserial correlation

Rank-Biserial correlation coefficient (r_{rb}) is a correlation coefficient used when one variable is dichotomous and the other ordinal. Mathematically, given a dichotomous variable Y and an ordinal variable X, the rank-biserial correlation coefficient is defined as

$$r_{bp} = \frac{2 * (M_1 - M_0)}{n}, \tag{3.7}$$

where M_1 is the mean value of the ordinal variable X when Y is 1, M_0 is the mean value of the ordinal variable X when Y is 0 and n is the total sample size. The interpretation of the results is still the same.

Phi coefficient φ

Phi coefficient φ (r_φ) is a measure of the association between two binary variables. This measure is a special case of Pearson's correlation for two binary variables. Mathematically, the phi coefficient is calculated as

$$\varphi = \sqrt{\frac{\chi^2}{n}}, \tag{3.8}$$

where n is the total number of observations. Unlike the other correlation coefficients, the values of the φ coefficient are between 0 and 1, as neither variable indicates order. Even so, the interpretation remains the same.

Goodman and Kruskal's lambda

Goodman and Kruskal's lambda is a measure of association for the contingency table of nominal variables. This measure is based on modal probabilities: it measures the percentage improvement in the probability of the dependent variable X (row variable) given the value of the independent variable Y (column variable). It is calculated as follows:

- S: the sum of the highest number in each row.

- R: the total of the highest row.
- N: the sum of all cells.
- lamda: (S – R) / (N – R)

3.2 Supervised classification

This section will explain the methods used to construct classifiers that discriminate between PD patients and HC subjects based on non-motor symptoms and the identification of the most relevant features in the process.

First, feature selection algorithms have been used to identify the most relevant features to discriminate PD patients from HC subjects and, subsequently, several classifiers have been trained with the available data, multiple quality indicators have been measured and the results have been compared statistically.

3.2.1 Feature selection

Feature selection is the process of selecting a subset of features when developing a predictive model. This technique is used to simplify models and make them easier to interpret. It also reduces training time and reduces overfitting.

Feature selection algorithms can be considered as search techniques to create subsets of features. Depending on the evaluation metric, three groups can be distinguished:

- Filter: Filter methods use proxy measures instead of error rate. These measures include mutual information [113], pointwise mutual information [114], and relief-based algorithms [115], among others.
- Wrapper: Wrapper methods use predictive models: for each subset the model is trained and the error rate of the model is obtained. They have a high computational cost, as a model has to be trained for each subset.
- Embedded: Embedded methods are techniques that are performed as part of the model building process. They can be prediction algorithms that have variable selection implemented, such as Random Forest, or LASSO methods with the L1 penalty to build a linear model.

Filter methods have a lower computational cost, although they are not tuned to a specific type of prediction model [116], so they tend to give worse prediction performance. However, as the subset of features are not bound to a prediction model, it is more useful for establishing the relationship between features. On the other hand, embedded methods are between filter and wrapper methods in terms of computational complexity, although they are also model-dependent. As several classification models are used in the project, filter methods have been selected.

Within the filter methods there are two subtypes: univariate and multivariate. The univariate method treats each variable independently, i.e. they evaluate the features according to certain criteria (e.g. euclidean distance to the class) and then select the best classified features. Whereas the methods in second subtype take groups of features into account when evaluating the performance. It has been decided to use multivariate methods, as they are able to deal with redundant, duplicated and correlated features.

Correlated Feature Selection

Correlated Feature Selection (CFS) is a correlation-based heuristic evaluation function that classifies features. This method was developed by Hall and Smith [117]. This function searches for subsets that are correlated with the class but independent of each other. The algorithm assumes that features that are irrelevant have a low correlation with the class, so they do not have to be included in the subsets. In addition, they examine excessive features, as these are often correlated with one of the other attributes. In this project we apply a forward best first search as a search heuristic with a stopping criterion of five consecutive fully expanded non-improving subsets.

In order to evaluate the subset S of k features, the following formula is used:

$$Merit_s = \frac{k \cdot \overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{3.9}$$

where $\overline{r_{cf}}$ is the average correlation value between the class and the features, and $\overline{r_{ff}}$ is the average correlation value between all pair of features.

3.2.2 Classifiers

In order to classify between PD patients and HC subjects we have used various machine learning techniques. The aim of machine learning (ML) is to develop techniques that allow computers to learn. Learning is considered to be the skill or knowledge that is acquired from experience [118]. Mitchell [119] provided a formal definition for the term: “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E”. In this case, the task is to make a binary classification. On the other hand, several performance measures are used in this project, which are explained in Section 3.2.4.

The baseline database has several predictor variables and a single class variable, so that supervised learning can be performed: the machine learning model learns to make predictions through a training process, correcting itself when the prediction is wrong. To be able to do this process, the data has to be divided into two parts, on the one hand the training data, with which the model is trained and corrected and, on the other hand, the test data, to be able to measure the level of accuracy achieved with the training data. This division is explained in Section 3.2.3.

The machine learning algorithms selected correspond to different types of groups. Algorithm groupings are made according to the similarity of their functions, which is why some of the algorithms can belong to more than one group. This section will explain the groups and algorithms used in the project:

- Instance-based Algorithms. These algorithms are based on the training data, predicting the new data using a similarity measure. Therefore, they are memory-based learning methods. The algorithms used in the project are K-Nearest Neighbors and Support Vector Machine.

- Decision Tree Algorithms. Using training data, these types of algorithms build a tree structure to make a prediction. The Decision Tree and Consolidated Tree Construction algorithms have been used in the project.
- Bayesian Algorithms. These algorithms apply Bayes theorem to perform the classification, e.g. Naive Bayes.
- Artificial Neural Network Algorithms. These algorithms are inspired by biological neural networks. The Multi-Layer Perceptron algorithm has been used in this project.
- Ensemble Algorithms. In these algorithms, multiple weaker models are trained independently and then combined to make a single prediction. For example, Random Forest combines several Decision Trees.
- Rule-based algorithm. These algorithms extract IF-THEN rules from the training data. The algorithm used is called RIPPER.

K-Nearest Neighbors (KNN)

KNN [120] is a non-parametric classification method, i.e. it assumes nothing about the underlying data. The algorithm calculates the distance of a new observation to the training observations, selects the K closest observations and assigns the new observation to the class to which most of the K observations belong.

Support Vector Machine (SVM)

SVM [121] are a set of supervised learning algorithms that construct a hyperplane or set of hyperplanes in a very high dimensional space to separate classes. In order to do this, the algorithm uses an optimisation process in which it finds the samples in the training data that are closest to the hyperplane that best separates the classes. These samples are called support vectors, therefore the name of the algorithm. In most cases it is impossible to separate the data correctly by a straight line, so the algorithm projects the data into a higher dimensional space. In order to solve this non-linear problem, kernels are used, which control the projection and the degree of flexibility in the separation of the classes [121].

Decision Tree (DT)

DT creates diagrams of logical constructs, which represent and categorize a series of conditions that occur in succession. In these tree structures, the leaves represent class labels and the branches represent the conjunctions of features leading to those class labels.

In this project the C4.5 algorithm [122], normally used for classification, has been used. This algorithm creates the decision trees from the training data. At each node (division) it chooses the attribute that most effectively divides the set of samples into subsets enriched in one class or another. Its criterion is the normalized one for information gain.

Consolidated Tree Construction (CTC)

CTC [123] is an algorithm designed to solve a class imbalance problem. It is based on the C4.5 algorithm, but uses a set of samples to build a single tree. To decide which feature to use in the partitioning, agreement is reached between the different sets of samples using a voting procedure. This procedure can be a standard voting, weighted voting or other strategies.

Naive Bayes (NB)

NB [124] is a probabilistic classifier based on applying Bayes theorem [125]. It is called *naive* because it assumes that a feature is independent of any other feature given the class variable. If we have n independent variables $(F_1, \cdots, F_n)$ and a class variable C, the probability using Bayes' theorem is the following:

$$p(C|F_1, \cdots, F_n) = \frac{p(C)p(F_1, \cdots, F_n|C)}{p(F_1, \cdots, F_n)}. \tag{3.10}$$

Multi-Layer Perceptron (MLP)

MLP [126] is an artificial neural network consisting of 3 types of layers: input layer, hidden layer and output layer. The layers are linked together and are composed of a set of nodes. Except for the nodes in the input layer, each node is a neuron using a non-linear activation function [127]. This network is trained using the *backpropagation* technique [128] and can distinguish data that is not linearly separable [129].

Random Forest (RF)

RF [130] builds a set of decision trees [131] in a parallel fashion, each trained on a slightly different sample generated by *bootstrapping* [132] and limiting randomly the features to be used in each tree and split. To predict a new observation, all decision trees are used independently and the new observation is assigned to the most common class.

Repeated Incremental Pruning to Produce Error Reduction (RIPPER)

RIPPER [133] algorithm induces a series of rules based on the training data. It uses the “separate and conquer" method, which adds conditions to a rule until it correctly classifies a subset of the data. Like decision trees, this algorithm also uses the information gain criterion to identify the best separation feature. The rule set is optimized using a series of heuristics.

3.2.3 Model validation

To make a better estimation, the models are validated using the 10-Fold cross-validation technique (10-Fold CV) [134]. The 10-Fold CV randomly divides the original sample into 10 sub-samples. One of the nine subsamples is used to test the model, while the remaining nine subsamples are used to train the model. This process is repeated 10 times for each of the 10 subsamples. Thus, 10 results are obtained and then averaged to evaluate the performance of the classifier. The same seed will be used in all classifiers, i.e. all classifiers will use the same 10 subsamples.

3.2.4 Performance measures

The performance of the models is calculated using evaluation metrics. These metrics are based on a confusion matrix, which captures the association between predictions and actual classes. The values of the confusion matrix can be seen in Table 3.3.

The metrics that will be calculated from this table are the following ones:

- Accuracy [135] shows how many of the samples have been correctly predicted. It is

	Predicted class	
Real class	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Table 3.3: Confusion matrix for a two-class problem. TP is the number of correct predictions of positive instances (true positive), FN is the number of incorrect predictions of negative instances (false negative), FP is the number of incorrect predictions of positive instances (false positive) and TN is the number of correct predictions of negative instances (true negative).

calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision [135] indicates the proportion of positive identifications that are correct. It is calculated as follows:

$$Precision = \frac{TP}{TP + FP}$$

- Recall [135] tells us the proportion of real positives correctly identified. It is calculated as follows:

$$Recall = \frac{TP}{TP + FN}$$

- F-Score [135] is a measure of the fidelity of a model, calculated from the harmonic mean between Precision and Recall. It is calculated as follows:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

- Area Under ROC curve [135] is a graphical representation of Sensitivity versus Specificity for every possible cut-off. Sensitivity is another name for the Recall metric and Specificity is calculated as follows:

$$Specifity = \frac{TN}{TN + FP}$$

  Area Under ROC curve is interpreted as follows:

  - $[0.5]$: Equitable to a random model.
  - $[0.5, 0.6)$: Bad model.
  - $[0.6, 0.75)$: Regular model.
  - $[0.75, 0.9)$: Good model.
  - $[0.9, 0.97)$: Very good model.
  - $[0.97, 1)$: Excellent model.

3.2.5 Statistical tests

In machine learning, several algorithms are often trained and compared in order to decide which one is more adequate for our objective. In order to compare them, it is not enough to observe the means of the different folds, since there are variations between them. One possibility is to use null hypothesis significance tests, although these are not very appropriate, since they do not calculate that one classifier is more accurate than another, but rather calculate the probability of obtaining the observed difference between the classifiers, assuming that the null hypothesis of equivalence is true.

Another way is to use a Bayesian analysis based on Bayesian estimation ([136, 137]). The Bayesian approach is based on the subjective interpretation of probability, which considers probability as a degree of belief with respect to uncertainty.

A parameter is seen as a random variable to which, prior to sampling evidence, an a prior probability distribution is assigned, based on a certain degree of belief regarding random behaviour. When the sampling evidence is obtained, the a prior distribution is modified and then an a posterior probability distribution emerges.

In this project the accuracies of all pairs of classifiers will be compared, taking into account the cross-validation of 10 folds with 10 repetitions. The same seed has been maintained in these 100 samples, i.e. the partitions used for every algorithm are identical. In order to make this comparison, we will perform the correlated Bayesian test proposed by Corani and Benavoli [138].

Bayesian correlated t-test

The test takes into account that cross-validation on a single database has correlations between training sets, based on the following generative model of the data:

$$x_{nx1} = 1_{nx1}\mu + v_{nx1}, \tag{3.11}$$

where x_{nx1} is the vector of accuracy differences„ 1_{nx1} is a vector of ones, μ is the parameter of interest and $v \sim MVN(0, \sum_{nxn})$ is a multivariate normal noise with zero mean and covariance matrix $\sum_{nxn}$. More details can be found in the article [139], Section 3.

The posterior distribution can be used to evaluate the probability of one of the algorithms being better than the other or of the two algorithms being “practically equivalent". To do this, we first have to define that two classifiers are practically equivalent if their mean difference of accuracies is less than a certain value (1% in our case), creating a region of practical equivalence (rope) [140] with the interval $[-0.01, 0.01]$. Once the rope is defined, the probabilities can be calculated from the posterior:

- P(left): the integral of the posterior in the interval $(-\infty, -0.01)$, namely the posterior probability that the mean difference in accuracies is practically negative.
- P(rope): the integral of the posterior in the interval $[-0.01, 0.01]$, namely the posterior probability that the two classifiers are practically equivalent.
- P(right): the integral of the posterior in the interval $(0.01, \infty)$, namely the posterior probability that the mean difference of the accuracies is practically positive.

CHAPTER 4

Implementation

This chapter presents the software and the procedure that has been carried out in the project. The work has been implemented using two main tools: on the one hand, Weka has been used for the execution of the algorithms together with the obtention of the metrics, since this tool allows the implementation of these algorithms in a simple and fast way; on the other hand, for everything else, such as for the analysis of the data or the statistical tests, the Python language has been used through the application of Jupyter Notebook.

Weka is a machine learning software implemented in Java language. It contains tools to perform data preparation, classification, regression, clustering, association rule mining and visualization. The interfaces that have been used are Weka Experimenter and Weka Explorer.

Jupyter Notebook is a web application that allows to create documents containing code, equations, visualizations and narrative text. Its uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization and machine learning, among others.

4.1 Database versions

The realization of the database versions has not been a trivial work. First of all, the versions of Andoni Angulo Celada's work were obtained, which were created based on clinical data that was available from PD patients and controls from the studies of our colleagues from Biocruces Bizkaia HRI. Subsequently, some variables were added, such as the medical questions of the SCOPA-AUT test (variables SCAU26A, SCAU26B and SCAU26C) and the variable related to the subject's status in the study was eliminated. Subsequently, some of the variables, such as SCAU_total, were updated.

The variable MoCA_total was also removed from the Individual and intermediates databases. This variable comes from the PPMI database, called MCATOT. This is why this variable was initially found in all the databases, but in order to have a concordance in the data set, it was only included in the Total version of the database. Nevertheless, before doing this process, it was verified that it did not affect the results of the data very much.

Finally, some names were changed, such as the one mentioned above, i.e. MoCA_total instead of MCATOT. We also modified some intermediate variables that only had a single individual variable, such as MCAABSTR. Therefore, the names of all the databases were unified.

4.2 Data description and preprocess

First of all, the Python language has been used to analyze and visualize the variables of Intermediate 2 and Total versions. *pandas* and *numpy* libraries have been used to work with the data. For visualization, *seaborn* and *matplotlib* libraries have been used. The first one for box plots and histograms, the second for pie charts and bar charts.

Then, *ttest_ind* function from *scipy* library has been used to calculate the p-value of the difference between classes. For this purpose, the examples in the Total database have been divided into two groups: a subset of subjects with PD and another subset of subjects with HC.

The code to compute the correlations has been implemented in Python as well. The variables used correspond to the Intermediate 2 and Total versions. In addition, the ordinal variables have been transformed using one-hot encoding: GENDER is transformed into GENDER0, GENDER1 and GENDER2, where the number indicates the corresponding category. The same happens with the HANDED variable, which has become HANDED1, HANDED2 and HANDED3. As for the Class variable, it has taken a value of 0 if the subject is HC and a value of 1 if the subject is PD. All other variables are continuous and no transformation has been applied to them.

On the one hand, the one-hot encoding has been implemented using *get_dummies* function of *pandas*. On the other hand, to calculate the Pearson coefficient correlation between the variables, the *corr* function of *pandas* package has been used. For Kendall's τ_b and Point-biserial, the *stats* library has been used and, for Rank-biserial and Goodman and Kruskal's lambda, a function has been created. The *crosstab* function of the *pandas* package has been used to create the contingency table for Goodman and Kruskal's lambda. As the φ coefficient has the same value as Pearson coefficient, it has been calculated using the *corr* function as well. Finally, for the visualisation, the *heatmap* function of the *seasborn* library has been used.

4.3 Feature selection

In this document, the final version of the feature selection is explained, although a more in-depth study was carried out. First, the CFS algorithm was run in Python and Weka. Different results were obtained through these algorithms, since they use different search heuristics to find the most optimal subsets. We also used the Fast Correlation Based Filter (FCBF) algorithm in Weka. FCBC [141] is a method based on information theory for feature selection. These methods were run on the entire database and it was found that depending on the algorithm or heuristic, different results were obtained.

In order to understand which of the algorithms provided a better subset of data, a 10-Fold CV was run, which provided in percentage which variables are more relevant. In this way it has been possible to quantify the relevance of the variables. The percentages of the algorithms in Weka were smaller, so the CFS algorithm implemented in Python obtained more stable subsets. For this reason this method has been used in the project.

CFS has been performed on all four versions of the database. Once the groups created for each fold are saved, the number of times each variable has been selected by CFS has been counted. For example, in the Individual version, the AGE variable is present in 7 of the 10 groups, i.e. 70 percent. In this way, the variables have been considerably reduced, since many of the variables do not appear in any subgroup.

Once the feature selection variables are obtained, it has been observed that some of them have appeared in all the folds, while others only in one. Therefore, a threshold has been set in order to further reduce the number of characteristics and only use those most relevant to the class. For this purpose, the threshold has been set at 50%, i.e. those characteristics that have appeared in at least half of the folds have been selected.

To implement the CFS in Python an implementation in github[1] has been used. To make the division of 10-Folds the function *KFold* of the library *sklearn* has been used. The implementation of CFS in Weka has been done using the function *CfsSubsetEval* and FCBF using the function *FCBFSearch*.

Multivariate feature selection showed that the AGE variable was always present. According to the PPMI study, subjects were selected taking into account age, so it should not be an important characteristic by itself. This is why it was decided to perform a ranking of the variables using two univariate feature selection techniques: Information Gain and Chi-Square Test. The implementation of both univariate filters was performed in Weka. The InfoGainAttributeEval function was used for the Information Gain and the ChiSquaredAttributeEval function for the Chi-Square Test.

4.4 Classification

After obtaining the 8 versions of the databases, i.e. 4 complete and 4 after applying CFS, it was decided which validation method to use. A cross validation was chosen, since this method allows to have an overview of the database. In addition, leave-one-out was discarded, since its computational cost was very high, especially in the MLP algorithm. Therefore, 10 repetitions of 10-Fold CV have been performed for the training. In order to compare the algorithms, several metrics of the 100 repetitions have been obtained. These metrics are the following: Accuracy, F-Score, Recall, Precision and Area Under the ROC curve.

On the other hand, when choosing the algorithms, we tried to cover all known types of groups. We also wanted to add a comprehensible algorithms that took into account the class imbalance, such as the CTC. As all the algorithms were implemented in Weka, it was decided to use this tool, as it was simple, efficient and especially fast. The Weka Experimenter interface has been used to train eleven different algorithms, using default parameters.

For KNN, the IBk algorithm has been used. When constructing a KNN classifier, it is necessary to choose the value of K. In order to find the best value for our task, it is proposed to be either 3 or 5. The distance used between the samples has been the Euclidean distance,

[1] https://github.com/ZixiaoShen/Correlation-based-Feature-Selection

which is calculated by measuring the straight line between two points. That is why it is necessary to use the normalized data.

The SVM algorithm has been interpreted using SMO. The complexity constant C value is 1. In this project we have used the polynomial kernel, which represents over polynomials of the original variables the similarity of training samples in a feature space, allowing the learning of nonlinear models. This algorithm assumes that the data are in a standard range, usually from 0 to 1. Therefore, in this project we have used the normalized data to train the SVM.

The C4.5 algorithm has been implemented using J48. This implementation has the confidence threshold for pruning set at 0.25 and the minimum number of instances per leaf at 2. On the other hand, the CTC algorithm implemented by J48Consolidate uses the same parameters. In addition, the coverage value that the user wants to achieve with the generated set of samples is 99.

The NaiveBayes algorithm has been used to implement NB. This implementation uses the Gaussian Naive Bayes algorithm [142], which assumes that the probability of the features is Gaussian.

The algorithm used to implement MLP is MultilayerPerceptron. We have implemented three architectures with 50, 100 and 200 neurons in the hidden layer and we have trained them through 500 epochs. The data used in this algorithm are normalized, as non-scaled input variables may result in a slow or unstable learning process. The activation function used in all nodes has been sigmoid and it is trained using the technique of *ADAM* [143].

Finally, RF has been implemented using RandomForest. The number of DT used is 100, which have a number of instances per leaf of 1, with no maximum depth.

4.5 Statistical tests

In order to compare the algorithms with each other, the one with the best Accuracy has been chosen among all the versions of the databases. In addition, in the case of the MLP and KNN algorithms, the parameter that obtained the best Accuracy has been selected.

Once the database versions for each algorithm to be compared were selected, their accuracy was compared by the Bayesian correlated t-test. Therefore, we statistically quantified which algorithm was better between pairs or whether their accuracy was similar. For this purpose, the 100 accuracy values obtained in the 10-Fold CV of 10 repetitions were taken into account. These values are comparable because the same seed has been used, i.e. the same data partition has been used for training and testing for all models. The implementation of the Bayesian correlated t-test has been also done through a implementation on github[2] in the Python language.

4.6 Rule obtention

Three of the algorithms used are not black box algorithms, i.e., they give a series of explanations such as rules or decision trees regarding their classification. These algorithms are RIPPER, DT and CTC.

[2] https://github.com/BayesianTestsML/tutorial/

To simplify the explanation of the models, the minimum number of instances per leaves (parameter M in the algorithm) or rules (parameter N in the algorithm), which by default in the algorithms is a value of 2, has been modified. Three values have been tested: 10, 40 and 50 and the models have been trained with a single repetition of Leave-one-out. Leave-one-out is a particular case of cross-validation, where a single sample is left for testing while all the others are left for training. This technique has been used because the trees or rules used are built with the whole set of rules, i.e., using almost the whole set of samples, identical or very similar trees have been built and therefore the real performance can be better approximated.

Only the two extremes of the database versions have been used, i.e. the Individual and Total databases. This is because these are the databases that are used in clinical analysis today and, as one of the objectives is to obtain biomarkers, it has been found more convenient to extract the relevant rules using the questions or the complete groupings of the tests.

This implementation has been done with the Weka Explorer interface. Through this interface it is possible to get the tree and the rules generated with the whole database.

4.7 Gender analysis

The aim was to see if there is any bias in the metrics depending on the subjects' gender, that is, if there is a gender that is better classified or if both genders are equally well classified. Remind that one third of the subjects are female and the rest are male.

In order to perform this analysis, the algorithms that best discriminated between PD and HC patients have been selected. These algorithms are SVM trained with the Individual + CFS database, RF trained with the Total database and MLP200 trained with the Total + CFS database.

The Weka Explorer interface has been used for the implementation. Through this interface, besides getting the desired metrics, it is possible to see the classification of each sample, i.e., which is the real class, which is the predicted class and some other attribute of the sample, in this case the gender.

Once this information is obtained, using the Python language, the data has been divided according to gender and then the desired metrics were obtained. To do this we have used the libraries of *pandas*, *numpy* and *sklearn*.

CHAPTER 5

Results

This section explains the obtained results. First, a brief analysis of the data will be shown, which includes the most relevant information. The details of the data and their corresponding graphs are included in the appendices. Subsequently, the selection of characteristics made and the results obtained will be explained. Finally, the results of the classifiers, their statistical comparison, the rules obtained from some of the algorithms and the gender analysis will be explained.

5.1 Data

5.1.1 Description

This section shows a brief description of the data, although a more detailed description of each variable, together with its corresponding graphs, can be found in Appendix B. As mentioned above, the analysis has been applied only to the Intermediate 2 and Total versions of the database.

Table 5.1 shows the mean and standard deviation of the quantitative variables in the Total version, divided by classes, as well as their p-values for the difference between the classes. Some variables have significant differences between the classes (< 0.001): GDS_total, SCAU_total, SDMTOTAL, UPSIT_total, MoCA_total, HVLTRT_total, HVLTRDLY and HVLTREC.

Variables	HC	PD	p-value
EDUCYRS	16.02 ± 2.89	15.51 ± 3.09	0.04
AGE	61.29 ± 11.17	62.03 ± 9.75	0.42
ESS_total	5.62 ± 3.40	6.09 ± 3.74	0.12
GDS_total	1.30 ± 2.09	2.46 ± 2.64	< 0.001
SCAU_total	7.44 ± 5.03	11.97 ± 7.51	< 0.001
SDMTOTAL	46.75 ± 10.50	41.16 ± 10.05	< 0.001
Benton_total	13.13 ± 1.97	12.78 ± 2.16	0.04
UPSIT_total	34.01 ± 4.84	23.43 ± 8.59	< 0.001
MoCA_total	28.22 ± 1.11	27.12 ± 2.34	< 0.001
HVLTRT_total	26.01 ± 4.48	24.44 ± 4.90	< 0.001
HVLTRDLY	9.26 ± 2.32	8.33 ± 2.57	< 0.001
HVLTREC	11.47 ± 0.83	11.13 ± 1.34	< 0.001

Table 5.1: Mean and standard deviation of the variables by class, and the p-value of the difference between classes. These variables are from the Total version.

5.1.2 Correlation

This section discusses the variables that are most correlated with each other. This test was performed taking into account all subjects, but also dividing the subjects among the classes. Appendix C shows all correlations using a heat map, as well as a more detailed analysis.

The analysis shows that for the SCOPA-AUT questionnaire, the variables are more related to each other than to the others. In the case of the MoCA test, on the other hand, the relationship is not so clear, the results of the questions are more heterogeneous. As for the Total version, there is no high correlation of these questionnaires with the other variables. In all cases there is a high correlation between HVLTRT_total and HVLTRDLY.

In relation to the class the most correlated variable in the Intermediate 2 version is UPSIT_total, with a negative correlation of -0.53, therefore the class is highly related to the smell. In addition, the variable SCAU_gastroint also stands out, with a positive correlation of 0.35. It seems that the information from a single variable does not provide enough information to be able to discriminate classes. In the Total version, it is repeated that the most correlated variable with the class is UPSIT_total. Focusing on positive values, the most correlated variable is SCAU_total (0.28), most likely due to the influence of SCAU_gastroint.

If we take into account the correlations using only HC subjects, there is a correlation between the AGE and SDMTOTAL variables, in addition to those mentioned above. In the case of only taking into account the PD subjects, no new correlations stand out.

5.2 Detection of Parkinson's disease

5.2.1 Feature Selection

This section explains the results obtained in the feature selection process. The aim is to find the most relevant features to discriminate between both classes, but that are not correlated between them, to avoid redundancy in the information. Thus, thanks to this process, not only will we be able to train the algorithms more efficiently, but we will also find out which subset of features is the most suitable for differentiating PD patients.

In Table 5.2 the features selected in each version of the database are shown. In addition, the last row shows the number of features in each version and and how much it has been reduced after applying the feature selection. For example, in the Individual database there are 106 independent variables and once CFS was applied it was reduced to 15 variables.

Total (7)	**Inter. 2 (7)**	**Inter. 1 (10)**	**Individual (15)**
HANDED			
AGE	AGE	AGE	AGE
GDS_total	–	–	GDSAFRAD, GDSENRGY
SCAU_total	SCAU_gastroint	SCAU_gastroint	SCAU1, SCAU2, SCAU5, SCAU6
			SCAU9
	SCAU_cardiovascular	SCAU_cardiovascular	SCAU15
	SCAU26B	SCAU26B	
UPSIT_total	UPSIT_total	UPSITBK1..4	UPSITBK1..4
MoCA_total		moca_naming	MCARHINO
	MCAABSTR	MCAABSTR	
	moca_recall		MCAREC2
HVLTREC			
14 -> 7	28 -> 7	52 -> 10	106 -> 15

Table 5.2: Summary of the independent variables used in each of the databases after feature selection. The last row shows the number of independent variables in the original database and after reduction by CFS.

The variable AGE appears in all versions, it seems that this variable provides additional information to another variable. It is noteworthy that the variables HANDED and HVLTREC only appear in the Total version, this may be due to the fact that CFS has less variables to select in this database. As for the GDS questionnaire, it can be observed that it is selected in both the Total and Individual versions, but not in the intermediate versions. In addition, the Individual variables do not correspond to the items in the GDS_6 variable, so it seems that the intermediate variable of GDS that we created based on previous publications is missing some important information.

The SCOPA-AUT questionnaire appears in all versions. Some of the questions on gastrointestinal problems and one question on cardiovascular problems are selected. As for variable SCAU26B, which asks about medication for urinary problems, it appears in the intermediate versions, while question SCAU9 ("In the past month, have you had involuntary loss of urine?") appears in the Individual version. These two variables are not correlated (0.00069), although both provide information related to urinary problems.

With regard to the UPSIT questionnaire, all variables appear in all databases. The correlation analysis shows that this questionnaire is correlated with the Class variable, so it is logical that the CFS method always chooses these variables.

Finally, the MoCA questionnaire is also present in all versions. It is noteworthy how the heuristic has chosen the variable MCAABSTR in the intermediate versions but not in the Individual one. It is also surprising that moca_recall is in the Intermediate 2 version and moca_naming in Intermediate 1, while both have an item variable that are comprised of variable in the Individual version.

The rankings generated by the Information Gain and Chi-Square Test methods show that most of the variables that appeared are at the top of the lists, although the AGE variable is not among them.

5.2.2 Classification

This section explains the classification results obtained. In order to get a general idea of the performance of the algorithms, the averages of the 10 repetitions of the 10-Fold CVs of the metrics have been calculated.

In the metrics, it has been taken into account that it is better to decrease the FN as much as possible. This is because patients with suspected PD are monitored for an average of two years before they are diagnosed. This is why a high Recall value is preferred to Precision.

Table 5.3 shows the results of the Individual database version with the complete database and after applying the CFS (Individual+CFS). It seems that feature reduction improves the performance of all algorithms except RF. If we look to the Accuracy, F-Score and Recall metrics the best classifier with the Individual database is RF, while after applying CFS it is SVM. RF algorithm has its own internal feature selection, so it works optimally even if there are many independent variables, 106 in this case. If we look at the Precision metric, in both cases the best algorithm is NB. Finally, in terms of AUC, the best classifier with the Individual database is again RF, but after applying CFS it is NB.

It is noteworthy that the NB presents good AUC, with a comparably small Accuracy. This is because the AUC analysis shows how well the positive class samples can be separated from the other class, i.e., how well the algorithm classifies PD samples, while the Accuracy indicates the actual performance of the algorithm.

Other algorithms that perform well on this database, in addition to the three mentioned above, are RIPPER, DT (when CFS is applied) and MLPs, especially with 200 neurons in the hidden layer. These algorithms have an Accuracy greater than 80%, F-Score greater than 0.85 and an AUC in the range of [0.75, 0.9).

Table 5.4 shows the results obtained in the Intermediate 1 version of the database with the complete database and after applying the CFS (Intermediate 1+CFS). Some of the algorithms improved the results compared to the Individual database, especially the NB (with a difference in Accuracy of 3.13). However, after applying CFS worse Accuracies were obtained than with the Individual+CFS (except MLPs and KNNs).

On the same database, better results were achieves after applying CFS, except for RF. Therefore, feature selection improved the performance of the algorithms. Regarding which algorithms obtain the best results, the same classifiers are repeated again: the best classifier in all metrics except Recall is RF, although after applying the CFS method it is SVM. Nevertheless, if we take into account the Precision metric (and AUC in the case of Intermediate 1+CFS), the best performing algorithm is NB. In addition to this, MLPs are also highlighted, being the best MLP100 in the Intermediate 1 and MLP200 after applying CFS. Finally, in Intermediate 1+CFS, RIPPER and KNNs also stand out, with KNN5 being

	Complete					CFS				
	Acc.	F1	Rec.	Prec.	AUC	Acc.	F1	Rec.	Prec.	AUC
CTC	77,95	0,83	0,78	0,90	0,77	79,39	0,84	0,78	0,92	0,81
RIPPER	81,34	0,87	0,87	0,87	0,78	81,79	0,87	0,88	0,87	0,78
DT	78,26	0,85	0,85	0,85	0,71	80,69	0,86	0,87	0,87	0,78
NB	76,94	0,82	0,74	**0,92**	0,87	81,47	0,86	0,78	**0,95**	**0,91**
SVM	82,61	0,88	0,88	0,88	0,79	**85,23**	**0,90**	**0,90**	0,90	0,82
RF	**83,41**	**0,89**	**0,92**	0,86	**0,90**	82,67	0,88	0,88	0,88	0,89
MLP50	80,51	0,86	0,85	0,87	0,88	81,68	0,87	0,86	0,88	0,88
MLP100	80,68	0,86	0,86	0,87	0,88	81,47	0,87	0,86	0,88	0,87
MLP200	80,74	0,86	0,86	0,87	0,88	82,11	0,87	0,87	0,88	0,88
KNN3	70,44	0,79	0,80	0,79	0,70	79,00	0,85	0,83	0,87	0,84
KNN5	72,61	0,81	0,82	0,80	0,73	79,89	0,86	0,83	0,88	0,86

Table 5.3: Metrics of the algorithms trained on the Individual database with the complete database and after applying the CFS. The best values for each metric are bolded in both cases.

the best.

	Complete					CFS				
	Acc.	F1	Rec.	Prec.	AUC	Acc.	F1	Rec.	Prec.	AUC
CTC	77,41	0,83	0,77	0,89	0,77	77,51	0,83	0,76	0,91	0,80
RIPPER	79,74	0,86	0,86	0,86	0,75	80,25	0,86	0,87	0,86	0,76
DT	77,83	0,84	0,84	0,85	0,72	79,45	0,85	0,85	0,86	0,77
NB	80,07	0,85	0,79	**0,92**	0,87	81,05	0,85	0,79	**0,94**	**0,89**
SVM	83,00	0,88	0,88	0,88	0,79	**83,92**	**0,89**	**0,88**	0,89	0,81
RF	**83,90**	**0,89**	**0,92**	0,87	**0,89**	81,77	0,87	0,88	0,87	0,87
MLP50	80,83	0,87	0,87	0,87	0,87	82,70	0,88	0,87	0,89	0,88
MLP100	80,80	0,87	0,87	0,87	0,87	82,58	0,88	0,87	0,89	0,88
MLP200	80,74	0,86	0,87	0,87	0,87	82,79	0,88	0,87	0,89	0,88
KNN3	69,73	0,79	0,80	0,78	0,70	80,22	0,86	0,85	0,87	0,81
KNN5	73,57	0,82	0,84	0,80	0,75	80,34	0,86	0,86	0,87	0,84

Table 5.4: Metrics of the algorithms trained on the Intermediate 1 database with the complete database and after applying the CFS. The best values for each metric are bolded in both cases.

Table 5.5 shows the results obtained in the Intermediate 2 database with the complete database and after applying the CFS (Intermediate 2+CFS). In general, this combination of features obtains better Accuracy than the combination created from the Intermediate 1 and Individual database, except for the NB algorithm and KNN5 algorithm. Regarding the comparison of the improvement with the CFS methods, as in the previous cases, the Accuracy improves in all algorithms except RF and in the case of KNN it improves considerably.

In the Intermediate 2 database the best performing algorithms were found to be SVM (if we look to the Accuracy and F-Score metrics) and RF (if we look to the Recall and AUC metrics). On the other hand, if we look to the Precision metric, the best classifier is CTC. Other notable algorithms are MLP and RIPPER, which although they are not the ones with the best average, in general they do not have bad values in their metrics. In the Intermediate 2 version the MLP50 works a little better, while in the Intermediate 2+CFS version the MLP100 is the best. If we look only at the CFS part in the table, other algorithms stand out in addition to those mentioned. These algorithms are RIPPER, DT and KNN, especially KNN5.

	Complete					CFS				
	Acc.	F1	Rec.	Prec.	AUC	Acc.	F1	Rec.	Prec.	AUC
CTC	79,68	0,85	0,79	**0,92**	0,80	80,31	0,85	0,78	**0,93**	0,83
RIPPER	81,38	0,87	0,88	0,86	0,77	82,23	0,87	0,88	0,88	0,79
DT	79,97	0,86	0,86	0,86	0,75	82,51	0,87	0,86	0,89	0,84
NB	79,69	0,82	0,76	0,90	0,84	79,59	0,84	0,79	0,92	0,87
SVM	**84,05**	**0,89**	0,88	0,89	0,81	**84,33**	**0,89**	0,88	0,90	0,82
RF	83,99	**0,89**	**0,92**	0,87	**0,89**	82,81	0,88	**0,89**	0,87	**0,89**
MLP50	81,63	0,87	0,88	0,87	0,87	83,13	0,88	**0,89**	0,88	0,88
MLP100	81,56	0,87	0,88	0,87	0,87	83,17	0,88	**0,89**	0,88	0,88
MLP200	81,60	0,87	0,88	0,87	0,87	82,98	0,88	0,88	0,88	0,88
KNN3	71,14	0,79	0,78	0,81	0,71	81,85	0,87	0,87	0,88	0,84
KNN5	71,79	0,80	0,80	0,8	0,74	82,05	0,87	0,87	0,88	0,86

Table 5.5: Metrics of the algorithms trained on the Intermediate 2 database with the complete database and after applying the CFS. The best values for each metric are bolded in both cases.

Table 5.6 shows the values of the algorithms trained with the Total version with the complete database and after applying the CFS (Total+CFS). If we compare it with the other versions, there is no clear improvement or worsening of the metrics, it always depends on which algorithm is taken into account. Most algorithms improve after applying CFS, but this difference is not as notorious as in previous versions of the databases.

Regarding the metrics in this table, it can be seen that the algorithm achieved with RF obtains very good results, followed by SVM. Once again, the Precision metric differs from the others, being the best classifier CTC in the Total version and NB y Total+CFS version. Other algorithms to be highlighted are RIPPER, DT and in the case of Total+CFS the MLPs, being MLP200 the best.

It has been observed that in most cases the Accuracy and F-Score metrics agree. Moreover, it is known that F-Score includes Precision and Recall. For this reason, from this point on, we have focused mainly on Accuracy, without forgetting the other metrics.

To have a more general overview of all databases together with the different algorithms, a bar chart has been created, which can be seen in Figure 5.1. It shows that SVM and RF have the best Accuracy and that feature selection helps in most cases, especially in KNN.

5.3 Statistical tests

In the previous section we have seen how, depending on the metric, some algorithms were found to be better than others. However, the Accuracy, F-Score and Recall results were in agreement. Therefore, this section has chosen the best models and compared them statistically using the Accuracy metric.

	Complete					CFS				
	Acc.	F1	Rec.	Prec.	AUC	Acc.	F1	Rec.	Prec.	AUC
CTC	80,24	0,85	0,81	**0,90**	0,80	81,04	0,86	0,82	0,91	0,82
RIPPER	81,66	0,87	0,88	0,87	0,78	81,30	0,87	0,87	0,87	0,77
DT	80,68	0,87	0,88	0,85	0,75	82,67	0,88	0,89	0,87	0,81
NB	77,58	0,83	0,78	0,89	0,85	79,47	0,84	0,79	**0,92**	0,87
SVM	83,29	0,88	0,88	0,89	0,80	83,7	0,88	0,87	0,90	0,81
RF	**84,65**	**0,89**	**0,91**	0,88	**0,90**	**84,31**	**0,89**	**0,91**	0,88	**0,90**
MLP50	80,42	0,86	0,88	0,85	0,86	83,19	0,88	0,88	0,88	0,88
MLP100	79,48	0,86	0,87	0,85	0,85	83,51	0,88	0,89	0,89	0,88
MLP200	79,33	0,86	0,87	0,85	0,85	83,58	0,89	0,89	0,88	0,88
KNN3	76,71	0,84	0,85	0,83	0,78	78,70	0,85	0,84	0,86	0,83
KNN5	77,44	0,84	0,86	0,83	0,81	80,69	0,86	0,85	0,88	0,86

Table 5.6: Metrics of the algorithms trained on the Total database with the complete database and after applying the CFS. The best values for each metric are bolded in both cases.

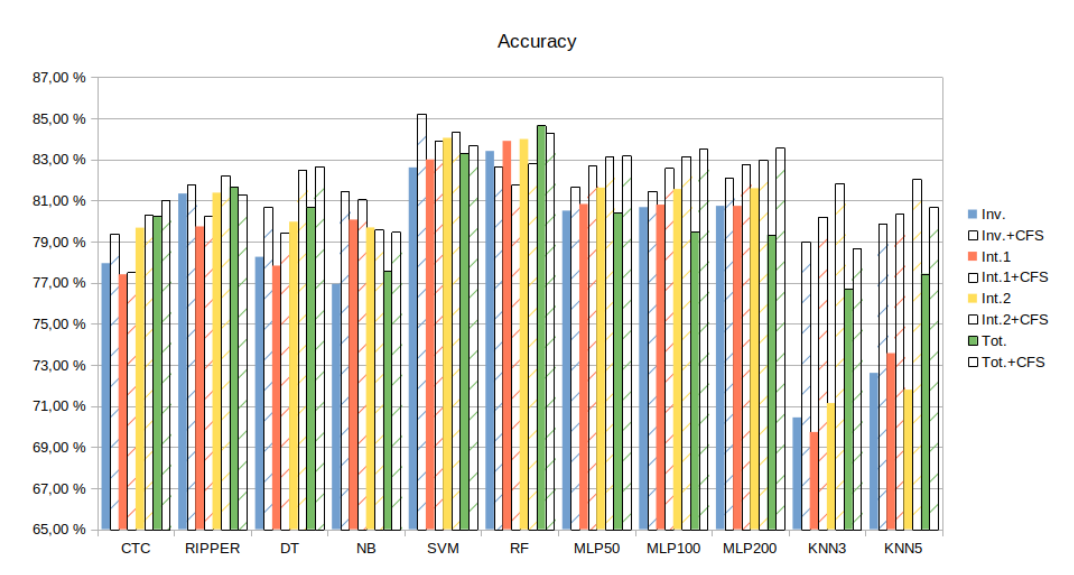


Figure 5.1: Bar chart of the Accuracy in the different versions of the database and the different algorithms.

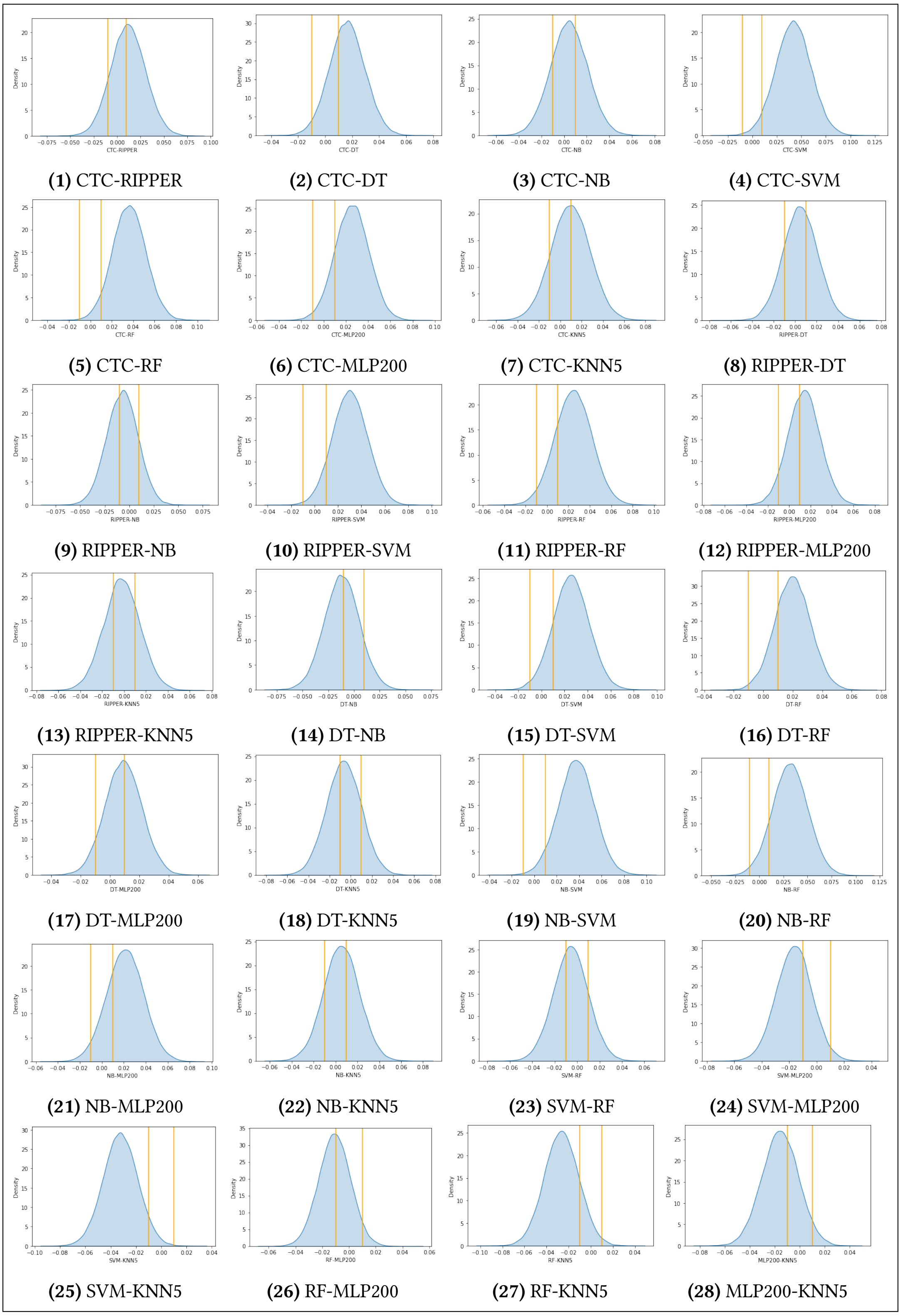
The databases with the best Accuracy in each model were selected. For CTC and DT the Total+CFS version was selected, for RF the Total version, for RIPPER the Intermediate 2+CFS version and for NB and SVM the Individual+CFS version. In addition, among the MLP models, the highest Accuracy has been selected taking into account the combination of database version and hyperparameter. In this case the MLP200 with the Total+CFS version was selected. The same criterion was used to select the KNN with the best Accuracy, being KNN5 with the Intermediate 2+CFS version.

In Table 5.7 these probabilities can be seen. In general, there is no single algorithm that is better than all the others. It can be said that with a high probability SVM and RF are better than CTC (P(CTC≪SVM)=0.961, P(CTC≪RF)=0.952). In addition, SVM is probably also better than NB (P(NB≪SVM)=0.956), although with RF it is not so clear (P(NB≪RF)=0.882).

	P(left)	P(rope)	P(right)
CTC-RIPPER	0.122	0.344	0.536
CTC-DT	0.023	0.295	0.682
CTC-NB	0.192	0.445	0.363
CTC-SVM	0.002	0.036	**0.961**
CTC-RF	0.001	0.045	**0.952**
CTC-MLP200	0.011	0.146	0.843
CTC-KNN5	0.141	0.369	0.490
RIPPER-DT	0.184	0.448	0.368
RIPPER-NB	0.434	0.427	0.139
RIPPER-SVM	0.004	0.083	0.913
RIPPER-RF	0.024	0.177	0.799
RIPPER-MLP200	0.059	0.344	0.597
RIPPER-KNN5	0.315	0.452	0.232
DT-NB	0.545	0.356	0.099
DT-SVM	0.012	0.147	0.841
DT-RF	0.008	0.201	0.791
DT-MLP200	0.064	0.464	0.472
DT-KNN5	0.419	0.424	0.157
NB-SVM	0.002	0.043	**0.956**
NB-RF	0.012	0.106	0.882
NB-MLP200	0.033	0.222	0.745
NB-KNN5	0.174	0.44	0.386
SVM-RF	0.392	0.451	0.157
SVM-MLP200	0.692	0.287	0.021
SVM-KNN5	0.947	0.052	0.001
RF-MLP200	0.525	0.43	0.042
RF-KNN5	0.855	0.135	0.010
MLP200-KNN5	0.653	0.307	0.040

Table 5.7: Probability that the algorithm on the left is better or both equal or the one on the right is better. Probabilities above 95% are bolded.

Another algorithm that is probably better than CTC is MLP200 (P(CTC≪MLP200)=0.843). If we compare these three algorithms with the rest we see that with high probability SVM is better than RIPPER (P(RIPPER≪SVM)=0.913), DT (P(DT≪SVM=0.841) and KNN5 (P(SVM≫KNN5)=0.947). On the other hand, RF also has a high probability of being better than KNN5 (P(RF≫KNN5)=0.855). If we compare these three algorithms, i.e. SVM, RF and MLP200, we cannot say statistically which one is better, nor that they are practically equivalent. In Figure 5.2 the distribution plots between all the selected pairs of algorithms can be seen.



(1) CTC-RIPPER **(2)** CTC-DT **(3)** CTC-NB **(4)** CTC-SVM

(5) CTC-RF **(6)** CTC-MLP200 **(7)** CTC-KNN5 **(8)** RIPPER-DT

(9) RIPPER-NB **(10)** RIPPER-SVM **(11)** RIPPER-RF **(12)** RIPPER-MLP200

(13) RIPPER-KNN5 **(14)** DT-NB **(15)** DT-SVM **(16)** DT-RF

(17) DT-MLP200 **(18)** DT-KNN5 **(19)** NB-SVM **(20)** NB-RF

(21) NB-MLP200 **(22)** NB-KNN5 **(23)** SVM-RF **(24)** SVM-MLP200

(25) SVM-KNN5 **(26)** RF-MLP200 **(27)** RF-KNN5 **(28)** MLP200-KNN5

Figure 5.2: Probability distribution graphics.

5.4 Rule obtention

This section explains the rules found using the CTC, DT and RIPPER algorithms. The aim is to find simple rules with good Accuracy.

Figure 5.3 shows the performance obtained by RIPPER changing the minimum number of instances covered by each rule. As it was expected, the more simple the created rules are, the worse the performance is, although it does not get much worse. We have used the Individual+CFS rules with M40, Total with M40 and Total+CFS with M40, because these have a good Accuracy and are simple rules.

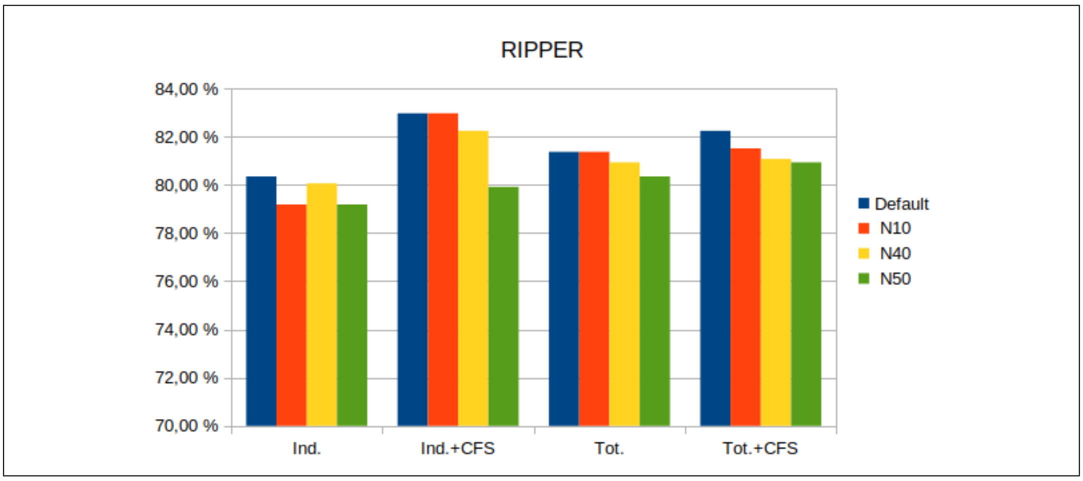


Figure 5.3: Accuracy of RIPPER by changing the minimum number of instances per rule (N).

Figure 5.4 shows how the Individual case improves considerably when it creates shallower trees. This is because deep trees do overfit the data, namely, they create trees that are able to classify all the training data, even the most specific samples, but do not generalize well to unseen instances.

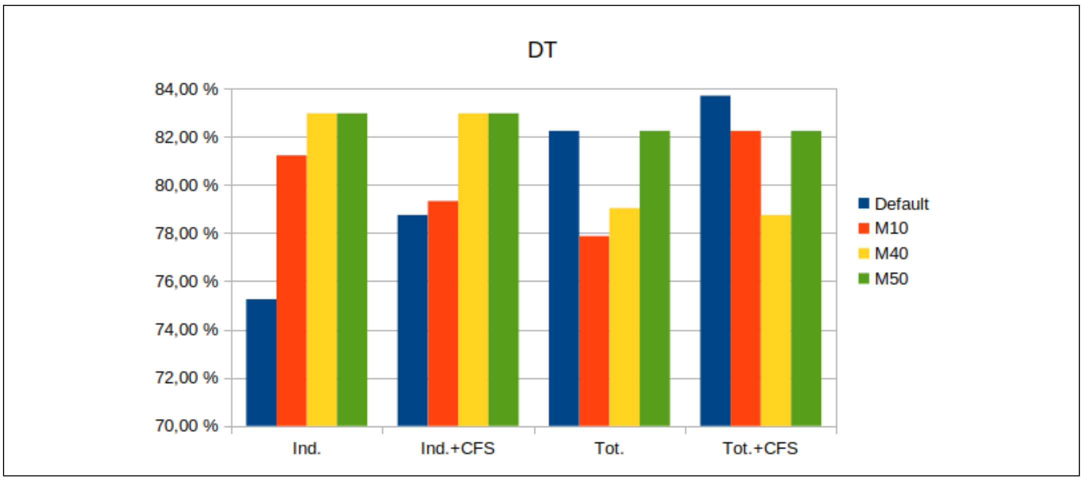


Figure 5.4: Accuracy of DT by changing the minimum number of instances per leaf (M).

When CFS is applied to the Individual database, less overfitting is achieved, although it continues to increase with parameter M. In the case of Total database it seems that the best fit are the parameters M=2 and M=50, although all have a performance from 77.87% to

82.24% (83.70% in the case of CFS). The trees converted to rules of Individual database with M40 and M50 were used along with Individual+CFS with M40 and M50, since all four gave the same tree, and Total with M50 and Total+CFS with M50, since they also gave the same tree.

Figure 5.5 shows the Accuracy of CTC when the minimum number of instances per leaf is changed. In this case the performance is much worse when M=40 or M=50. Moreover, the trees created with M=2 or M=10 are too complex to obtain simple rules. This is why no rules have been obtained from this algorithm.

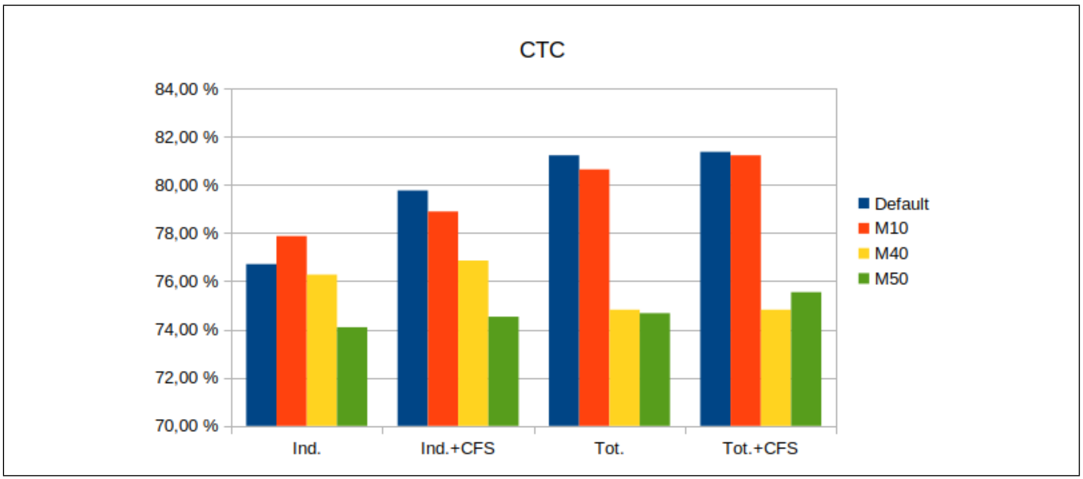


Figure 5.5: Accuracy of CTC by changing the minimum number of instances per leaf (M).

The rules in the Individual database are the following:

- UPSITBK1 < 7 or GDSENRGY $\neq 0 \rightarrow$ PD
- UPSITBK1 < 7 or UPSITBK3 ≤ 7 or SCAU2 $\neq 0 \rightarrow$ PD

The rules in the Total database are the following:

- UPSIT_total < 31 or SCAU_total $> 11 \rightarrow$ PD
- UPSIT_total < 30 or SCAU_total $> 10 \rightarrow$ PD
- UPSIT_total < 30 or SCAU_total $> 11 \rightarrow$ PD

The individual rules have the same first element, but differ in everything else. The first one corresponds to the RIPPER algorithm, with an Accuracy of 82.24%, while the second one to the DT with a very similar accuracy of 82.97%. As for the total rules, it seems clear that UPSIT_total and SCAU_total are the most relevant features although the rules slightly disagree in the particular thresholds. The first two rules correspond to RIPPER, with an Accuracy of 80.93% and 81.08% and the third to DT with an Accuracy of 82.24%.

5.5 Gender Study

Finally, it has been studied whether the best classifiers, i.e. SVM, ML200 and RF, achieve the same results for both men and women. For this purpose, Accuracy, Recall and Precision metrics have been obtained for each gender.

Table 5.8 shows the results of the metrics for each classifier. The MLP algorithm performs in a similar way for male and female samples. However, the SVM and RF algorithms achieve better Accuracy in the case of males (especially the RF algorithm). It should be noted that the improvement that RF achieves with males is through a better Precision, while SVM achieves a better Recall.

		Accuracy	Recall	Precision
SVM	Men	0.866	0.919	0.896
	Women	0.841	0.863	0.906
MLP200	Men	0.844	0.910	0.877
	Women	0.833	0.917	0.856
RF	Men	0.846	0.873	0.909
	Women	0.816	0.893	0.852

Table 5.8: Metrics by gender.

CHAPTER 6

Discussion

Although further studies are needed to refine the outcome of the work, considering the results of the feature selection process and classification, we believe that the result of this work is a promising step towards early and rapid detection of PD.

First, the descriptive study of the data suggests that some of the tests have a significant difference between the classes and that this information is not correlated with each other. However, the test that stands out the most, i.e., the one with the highest correlation with the class is the UPSIT test.

On the other hand, the results of the feature selection process suggest that GDS, SCOPA-AUT, UPSIT, MoCA and HVLTR are the most informative tests for discriminating PDs from HCs, while ESS, SDMT and BJLOT tests are probably more informative for other qualities of PDs. Also clinical variables such as age or dominant hand provide information to these tests to discriminate patients with Parkinson's disease.

On the other hand, by means of feature selection, better results are achieved in some of the algorithms. This is especially seen in the case of KNN, since the improvement is higher, although it also influences the other algorithms except RF. RF has its own internal feature selection.

There is no one version of the database that stands out from the others, the metric and the classifier must always be taken into account. Nevertheless, none of the best results have been obtained with the Intermediate 1 database, neither before nor after applying the CFS method, which suggests that the combination of features in Intermediate 2 is more appropriate. On the other hand, all the best algorithms, except RF, have got their best results with a feature reduced database version.

In addition, obtaining information from the different versions of the databases, it could be said that these tests do not have to be performed completely, but only some of their parts or some of the questions. In the case of the UPSIT test, it is always suggested to perform it completely. This reinforces the belief that a combination of clinical tests could be created, with a selected items from each tests, in order to detect PD.

All the classifiers selected for statistical comparison are good models, or very good models in the case of RF and NB, according to the AUC metric. However, according to the

Accuracy metric, which can be equated to the F-Score and Recall results, the best classifiers are RF, SVM and MLP200. There is no statistical evidence that any of these classifiers is better than the other, nor that they are equivalent. However, statistical comparison with other classifiers shows that SVM is better than all the others with high probability.

Previous studies have used Machine Learning techniques to discriminate PD patients from controls using PPMI data, combine imaging data with clinical data. These works have used SVM algorithm together some feature selection technique, such as Fisher discriminant ratio [96] or Random Forest [9]. However, Singh et al. [96] do not study early detection of the disease. Regarding Amoroso et al. [9], they show the brain regions mostly affected by the disease but they did not compare the classification performance of different algorithms as we did in the current work.

If we take into account the gender perspective, we find a mismatch in the case of SVM and RF. The first model has a higher percentage of false negatives in women, as opposed to the second, which has a higher percentage in men. Among the three mentioned classifiers, MLP200 is the one with the smallest gender bias, and the one with the smallest difference in false negatives. It is believed that by adjusting the hyperparameters of the MLP algorithm to our objective in a deeper way we could obtain an even better classifier to differentiate PD subjects from HC subjects, since this algorithm is the one that offers the greatest tuning opportunities.

Finally, we wanted to use classifiers with explaining capacity to obtain comprehensible rules applicable to the database. The rules obtained have an Accuracy higher than 80%, moreover, they are simple rules that just use the UPSIT, SCAU and GDS tests, further reducing the tests that would have to be applied. The rules obtained using the individual version of the database suggest that the UPSITBK1 test should be taken into account. As for the total database, a very good PD detection capacity was seen for the complete UPSIT and SCOPA-AUT tests.

CHAPTER 7

Conclusion and future work

In this work the problem of early detection of Parkinson's disease is studied, based on non-motor symptoms and exploring the best combination of items from clinical questionnaires and tests by means of Machine Learning techniques. The selection of variables using the CFS method has proven to be adequate because it considerably reduces the number of items, keeping or even increasing the discriminating capacity to differentiate PD patients from controls. It has been seen that the most important variables are related to the UPSIT test, which measures olfactory capacity. In addition, a way to group these items semantically is proposed. To test the effectiveness of the proposal, several algorithms are evaluated. The best results were obtained with the RF, SVM and MLP algorithms, with an Accuracy around 85%. The result was not only satisfactory because of the metrics achieved, but also because the number of variables could be reduced considerably, selecting the most relevant elements or grouping some of them semantically. It has also been seen that MLP algorithm had little bias regarding to the gender attribute. Finally, a simple rule capable of differentiating HC subjects from PD patients is proposed.

As future work, we plan to carry out a more in-depth study, on the one hand using the data provided by the Biocruces Bizkaia Health Research Institute and, on the other hand, to improve the classifiers to obtain better results. To this end, a search for the parameters that best fit the objective of differentiating between PD and HC patients is proposed. Also, future works require to test another series of machine learning algorithms or to use some deep learning models.

As for the gender bias, there are two ways of dealing with it. The first is to use specific techniques for training, while the second is to train two different algorithms, one with the men's data and the other with the women's data.

In the same way that gender is taken into account, it has been proposed to divide the database according to age, since it has been seen that age influences on symptoms. In other words, it is proposed to divide the database by age range, to extract the most relevant characteristics of each age range, and to train the algorithms with these data

Finally, it would also be interesting to perform a longitudinal study using non-motor variables to assess whether early motor symptoms can serve as a biomarker to predict disease progression.

Bibliography

[1] Gabriel Solana-Lavalle, Juan-Carlos Galán-Hernández, and Roberto Rosas-Romero. Automatic parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features. *Biocybernetics and Biomedical Engineering*, 40(1):505–516, 2020. See pages 1, 2.

[2] Arman Rahmim, Peng Huang, Nikolay Shenkov, Sima Fotouhi, Esmaeil Davoodi-Bojd, Lijun Lu, Zoltan Mari, Hamid Soltanian-Zadeh, and Vesna Sossi. Improved prediction of outcome in parkinson's disease using radiomics analysis of longitudinal dat spect images. *NeuroImage: Clinical*, 16:539–544, 2017. See pages 1, 10.

[3] Francisco PM Oliveira, Diogo Borges Faria, Durval C Costa, Miguel Castelo-Branco, and João Manuel RS Tavares. Extraction, selection and comparison of features for an effective automated computer-aided diagnosis of parkinson's disease based on [123 i] fp-cit spect images. *European journal of nuclear medicine and molecular imaging*, 45(6):1052–1062, 2018. See pages 1, 9.

[4] Zehra Karapinar Senturk. Early diagnosis of parkinson's disease using machine learning algorithms. *Medical hypotheses*, 138:109603, 2020. See pages 1, 2, and 11.

[5] Clayton R Pereira, Danilo R Pereira, Silke AT Weber, Christian Hook, Victor Hugo C de Albuquerque, and Joao P Papa. A survey on computer-assisted parkinson's disease diagnosis. *Artificial intelligence in medicine*, 95:48–63, 2019. See pages 1, 2.

[6] Koichiro Yasaka, Koji Kamagata, Takashi Ogawa, Taku Hatano, Haruka Takeshige-Amano, Kotaro Ogaki, Christina Andica, Hiroyuki Akai, Akira Kunimatsu, Wataru Uchida, et al. Parkinson's disease: deep learning with a parameter-weighted structural connectome matrix for diagnosis and neural circuit disorder investigation. *Neuroradiology*, pages 1–12, 2021. See pages 1, 9.

[7] Ane Murueta-Goyena, Ane Andikoetxea, Juan Carlos Gómez-Esteban, and Iñigo Gabilondo. Contribution of the gabaergic system to non-motor manifestations in premotor and early stages of parkinson's disease. *Frontiers in pharmacology*, 10:1294, 2019. See pages 1, 2, and 4.

[8] Diego Castillo-Barnes, Javier Ramírez, Fermín Segovia, Francisco J Martínez-Murcia, Diego Salas-Gonzalez, and Juan M Górriz. Robust ensemble classification methodology for i123-ioflupane spect images and multiple heterogeneous biomarkers in the diagnosis of parkinson's disease. *Frontiers in neuroinformatics*, 12:53, 2018. See pages 1, 10.

[9] Nicola Amoroso, Marianna La Rocca, Alfonso Monaco, Roberto Bellotti, and Sabina Tangaro. Complex networks reveal early mri markers of parkinson's disease. *Medical image analysis*, 48:12–24, 2018. See pages 2, 5, 10, and 46.

[10] SD Karamintziou, B Piallat, Stéphan Chabardès, Mircea Polosan, Olivier David, George L Tsirogiannis, NG Deligiannis, PG Stathis, George A Tagaris, EJ Boviatsis, et al. Design of a novel closed-loop deep brain stimulation system for parkinson's disease and obsessive-compulsive disorder. In *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 860–863. IEEE, 2015. See page 2.

[11] Andrew J Hughes, Susan E Daniel, Linda Kilford, and Andrew J Lees. Accuracy of clinical diagnosis of idiopathic parkinson's disease: a clinico-pathological study of 100 cases. *Journal of neurology, neurosurgery & psychiatry*, 55(3):181–184, 1992. See page 2.

[12] Thomas Foltynie, Carol Brayne, and Roger A Barker. The heterogeneity of idiopathic parkinson's disease. *Journal of neurology*, 249(2):138–145, 2002. See pages 2, 3.

[13] AJ Lees and Eileen Smith. Cognitive deficits in the early stages of parkinson's disease. *Brain*, 106(2):257–270, 1983. See page 3.

[14] AM Owen, M James, PN Leigh, BA Summers, CD Marsden, NP1 al Quinn, Klaus W Lange, and TW Robbins. Fronto-striatal cognitive deficits at different stages of parkinson's disease. *Brain*, 115(6):1727–1751, 1992. See page 3.

[15] RG Brown and CD Marsden. How common is dementia in parkinson's disease? *The Lancet*, 324(8414):1262–1265, 1984. See page 3.

[16] Patricia M Fitzgerald and Joseph Jankovic. Lower body parkinsonism: evidence for vascular etiology. *Movement disorders: official journal of the Movement Disorder Society*, 4(3):249–260, 1989. See page 3.

[17] ES Tolosa and J Santamaria. Parkinsonism and basal ganglia infarcts. *Neurology*, 34(11):1516–1516, 1984. See page 3.

[18] John Winikates and Joseph Jankovic. Clinical correlates of vascular parkinsonism. *Archives of neurology*, 56(1):98–102, 1999. See page 3.

[19] JCM Zijlmans, HOM Thijssen, OJM Vogels, HPH MD PhD Kremer, PJE Poels, HC Schoonderwaldt, JL Merx, MA Van't Hof, Th Thien, and MWIM Horstink. Mri in patients with suspected vascular parkinsonism. *Neurology*, 45(12):2183–2188, 1995. See page 3.

[20] Amos D Korczyn. Vascular parkinsonism—characteristics, pathogenesis and treatment. *Nature Reviews Neurology*, 11(6):319, 2015. See page 3.

[21] Hae-Won Shin and Sun Ju Chung. Drug-induced parkinsonism. *Journal of clinical neurology (Seoul, Korea)*, 8(1):15, 2012. See page 3.

[22] Ahmed A Moustafa, Srinivasa Chakravarthy, Joseph R Phillips, Ankur Gupta, Szabolcs Keri, Bertalan Polner, Michael J Frank, and Marjan Jahanshahi. Motor symptoms in parkinson's disease: A unified framework. *Neuroscience & Biobehavioral Reviews*, 68:727–740, 2016. See pages 3, 4.

[23] R Bhidayasiri. Differential diagnosis of common tremor syndromes. *Postgraduate medical journal*, 81(962):756–762, 2005. See page 3.

[24] Mary Ann Thenganatt and Elan D Louis. Distinguishing essential tremor from parkinson's disease: bedside tests and laboratory evaluations. *Expert review of neurotherapeutics*, 12(6):687–696, 2012. See page 3.

[25] Rick C Helmich, Mark Hallett, Günther Deuschl, Ivan Toni, and Bastiaan R Bloem. Cerebral causes and consequences of parkinsonian resting tremor: a tale of two circuits? *Brain*, 135(11):3206–3226, 2012. See page 3.

[26] Joseph Jankovic, M McDermott, J Carter, S Gauthier, C Goetz, L Golbe, S Huber, W Koller, C Olanow, I Shoulson, et al. Variable expression of parkinson's disease: A base-line analysis of the dat atop cohort. *Neurology*, 40(10):1529–1529, 1990. See page 3.

[27] Jong Moon Lee, Seong-Beom Koh, Sung Won Chae, Woo-Keun Seo, Ji Hyun Kim, Kyungmi Oh, Jong Sam Baik, Kun Woo Park, et al. Postural instability and cognitive dysfunction in early parkinson's disease. *Canadian journal of neurological sciences*, 39(4):473–482, 2012. See page 3.

[28] JE McLennan, K Nakano, HR Tyler, and RS Schwab. Micrographia in parkinson's disease. *Journal of the neurological sciences*, 15(2):141–152, 1972. See page 4.

[29] Aileen K Ho, Robert Iansek, Caterina Marigliani, John L Bradshaw, and Sandra Gates. Speech impairment in a large sample of patients with parkinson's disease. *Behavioural neurology*, 11(3):131–137, 1998. See page 4.

[30] Sabine Skodda, Wenke Visser, and Uwe Schlegel. Gender-related patterns of dysprosody in parkinson disease and correlation between speech variables and motor symptoms. *Journal of Voice*, 25(1):76–82, 2011. See page 4.

[31] Stewart J Fellows, J Noth, and M Schwarz. Precision grip and parkinson's disease. *Brain: a journal of neurology*, 121(9):1771–1784, 1998. See page 4.

[32] Diego Garcia-Borreguero, Oscar Larrosa, and Mauricio Bravo. Parkinson's disease and sleep. *Sleep medicine reviews*, 7(2):115–129, 2003. See page 4.

[33] Alex Iranzo, José Luis Molinuevo, Joan Santamaría, Mónica Serradell, María José Martí, Francesc Valldeoriola, and Eduard Tolosa. Rapid-eye-movement sleep behaviour disorder as an early marker for a neurodegenerative disorder: a descriptive study. *The Lancet Neurology*, 5(7):572–577, 2006. See page 4.

[34] Janice C Wong, Yanping Li, Michael A Schwarzschild, Alberto Ascherio, and Xiang Gao. Restless legs syndrome: an early clinical feature of parkinson disease in men. *Sleep*, 37(2):369–372, 2014. See page 4.

[35] RD Abbott, GW Ross, LR White, CM Tanner, KH Masaki, JS Nelson, JD Curb, and H Petrovitch. Excessive daytime sleepiness and subsequent development of parkinson disease. *Neurology*, 65(9):1442–1446, 2005. See page 4.

[36] MD Gjerstad, T Wentzel-Larsen, D Aarsland, and JP Larsen. Insomnia in parkinson's disease: frequency and progression over time. *Journal of Neurology, Neurosurgery & Psychiatry*, 78(5):476–479, 2007. See page 4.

[37] Mary A Carskadon, Edward D Brown, and William C Dement. Sleep fragmentation in the elderly: relationship to daytime sleep tendency. *Neurobiology of aging*, 3(4):321–327, 1982. See page 4.

[38] JF Pagel. Excessive daytime sleepiness. *American family physician*, 79(5):391–396, 2009. See page 4.

[39] Christopher J Earley. Restless legs syndrome. *New England Journal of Medicine*, 348(21):2103–2109, 2003. See page 4.

[40] Yves Dauvilliers, Carlos H Schenck, Ronald B Postuma, Alex Iranzo, Pierre-Herve Luppi, Giuseppe Plazzi, Jacques Montplaisir, and Bradley Boeve. Rem sleep behaviour disorder. *Nature reviews Disease primers*, 4(1):1–16, 2018. See page 4.

[41] Jack J Chen and Laura Marsh. Anxiety in parkinson's disease: identification and management. *Therapeutic advances in neurological disorders*, 7(1):52–59, 2014. See page 4.

[42] Jennifer SAM Reijnders, Uwe Ehrt, Wim EJ Weber, Dag Aarsland, and Albert FG Leentjens. A systematic review of prevalence studies of depression in parkinson's disease. *Movement disorders*, 23(2):183–189, 2008. See page 4.

[43] Thomas R Barber, Michael Lawton, Michal Rolinski, Samuel Evetts, Fahd Baig, Claudio Ruffmann, Aimie Gornall, Johannes C Klein, Christine Lo, Gary Dennis, et al. Prodromal parkinsonism and neurodegenerative risk stratification in rem sleep behavior disorder. *Sleep*, 40(8), 2017. See page 4.

[44] EL Jacob, NM Gatto, A Thompson, Y Bordelon, and B Ritz. Occurrence of depression and anxiety prior to parkinson's disease. *Parkinsonism & related disorders*, 16(9):576–581, 2010. See page 4.

[45] Richard L Doty. Olfactory dysfunction in parkinson disease. *Nature Reviews Neurology*, 8(6):329–339, 2012. See page 4.

[46] Stefan Brodoehl, Carsten Klingner, Gerd F Volk, Thomas Bitter, Otto W Witte, and Christoph Redecker. Decreased olfactory bulb volume in idiopathic parkinson's disease detected by 3.0-tesla magnetic resonance imaging. *Movement disorders*, 27(8):1019–1025, 2012. See page 4.

[47] Jia Li, Cheng-zhi Gu, Jian-bin Su, Lian-hai Zhu, Yong Zhou, Huai-yu Huang, and Chun-feng Liu. Changes in olfactory bulb volume in parkinson's disease: a systematic review and meta-analysis. *PLoS One*, 11(2):e0149286, 2016. See page 4.

[48] Nermin Tanik, Halil Ibrahim Serin, Asuman Celikbilek, Levent Ertugrul Inan, and Fatma Gundogdu. Associations of olfactory bulb and depth of olfactory sulcus with basal ganglia and hippocampus in patients with parkinson's disease. *Neuroscience letters*, 620:111–114, 2016. See page 4.

[49] Christoph Scherfler, Regina Esterhammer, Michael Nocker, Philipp Mahlknecht, Heike Stockner, Boris Warwitz, Sabine Spielberger, Bernadette Pinter, Eveline Donnemiller, Clemens Decristoforo, et al. Correlation of dopaminergic terminal dysfunction and microstructural abnormalities of the basal ganglia and the olfactory tract in parkinson's disease. *Brain*, 136(10):3028–3037, 2013. See page 4.

[50] Christoph Scherfler, Michael F Schocke, Klaus Seppi, Regina Esterhammer, Christian Brenneis, Werner Jaschke, Gregor K Wenning, and Werner Poewe. Voxel-wise analysis of diffusion weighted imaging reveals disruption of the olfactory tract in parkinson's disease. *Brain*, 129(2):538–542, 2006. See page 4.

[51] RKB Pearce, CH Hawkes, and SE Daniel. The anterior olfactory nucleus in parkinson's disease. *Movement disorders: official journal of the Movement Disorder Society*, 10(3):283–287, 1995. See page 4.

[52] Antony J Harding, Emily Stimson, Jasmine M Henderson, and Glenda M Halliday. Clinical correlates of selective pathology in the amygdala of patients with parkinson's disease. *Brain*, 125(11):2431–2445, 2002. See page 4.

[53] B Westermann, E Wattendorf, U Schwerdtfeger, A Husner, P Fuhr, O Gratzl, T Hummel, D Bilecen, and A Welge-Lüssen. Functional imaging of the cerebral olfactory system in patients with parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(1):19–24, 2008. See page 4.

[54] Nicolaas I Bohnen, Martijn LTM Müller, Vikas Kotagal, Robert A Koeppe, Michael A Kilbourn, Roger L Albin, and Kirk A Frey. Olfactory dysfunction, central cholinergic integrity and cognitive impairment in parkinson's disease. *Brain*, 133(6):1747–1754, 2010. See page 4.

[55] Ming-Ching Wen, Zheyu Xu, Zhonghao Lu, Ling Ling Chan, Eng King Tan, and Louis CS Tan. Microstructural network alterations of olfactory dysfunction in newly diagnosed parkinson's disease. *Scientific reports*, 7(1):1–9, 2017. See page 4.

[56] S Righi, MP Viggiano, M Paganini, S Ramat, and P Marini. Recognition of category-related visual stimuli in parkinson's disease: before and after pharmacological treatment. *Neuropsychologia*, 45(13):2931–2941, 2007. See page 4.

[57] Ana Marques, Kathy Dujardin, Muriel Boucart, Delphine Pins, Marie Delliaux, Luc Defebvre, Philippe Derambure, and Christelle Monaca. Rem sleep behaviour disorder and visuoperceptive dysfunction: a disorder of the ventral visual stream? *Journal of neurology*, 257(3):383–391, 2010. See page 4.

[58] Olga Veselá, Evžen Růžička, Robert Jech, Jan Roth, Kateřina Štěpánková, Petr Mečíř, Zuzana Solano, and Eva Preclíková. Colour discrimination impairment is not a reliable early marker of parkinson's disease. *Journal of neurology*, 248(11):975–978, 2001. See page 4.

[59] RB Postuma, JF Gagnon, M Vendette, and JY Montplaisir. Markers of neurodegeneration in idiopathic rapid eye movement sleep behaviour disorder and parkinson's disease. *Brain*, 132(12):3298–3307, 2009. See page 4.

[60] L Ferini-Strambi, MR Di Gioia, V Castronovo, A Oldani, M Zucconi, and SF Cappa. Neuropsychological assessment in idiopathic rem sleep behavior disorder (rbd): does the idiopathic form of rbd really exist? *Neurology*, 62(1):41–45, 2004. See page 4.

[61] Jean-François Gagnon, Mélanie Vendette, Ronald B Postuma, Catherine Desjardins, Jessica Massicotte-Marquez, Michel Panisset, and Jacques Montplaisir. Mild cognitive impairment in rapid eye movement sleep behavior disorder and parkinson's disease. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 66(1):39–47, 2009. See page 4.

[62] Dag Aarsland, K Brønnick, JP Larsen, OB Tysnes, G Alves, et al. Cognitive impairment in incident, untreated parkinson disease: the norwegian parkwest study. *Neurology*, 72(13):1121–1126, 2009. See page 4.

[63] Maria Livia Fantini, Elena Farini, Paola Ortelli, Marco Zucconi, Mauro Manconi, Stefano Cappa, and Luigi Ferini-Strambi. Longitudinal study of cognitive function in idiopathic rem sleep behavior disorder. *Sleep*, 34(5):619–625, 2011. See page 4.

[64] Seung-Hyun Kim, Ji-Hye Park, Yu Hwan Kim, and Seong-Beom Koh. Stereopsis in drug naive parkinson's disease patients. *Canadian journal of neurological sciences*, 38(2):299–302, 2011. See page 4.

[65] Kazumi Ota, Hiroshige Fujishiro, Koji Kasanuki, Daizo Kondo, Yuhei Chiba, Norio Murayama, Heii Arai, Kiyoshi Sato, and Eizo Iseki. Prediction of later clinical course by a specific glucose metabolic pattern in non-demented patients with probable rem sleep behavior disorder admitted to a memory clinic: a case study. *Psychiatry Research: Neuroimaging*, 248:151–158, 2016. See page 4.

[66] Per Svenningsson, Eric Westman, Clive Ballard, and Dag Aarsland. Cognitive impairment in patients with parkinson's disease: diagnosis, biomarkers, and treatment. *The Lancet Neurology*, 11(8):697–707, 2012. See page 4.

[67] Michele Terzaghi, Chiara Zucchella, Valter Rustioni, Elena Sinforiani, and Raffaele Manni. Cognitive performances and mild cognitive impairment in idiopathic rapid eye movement sleep behavior disorder: results of a longitudinal follow-up study. *Sleep*, 36(10):1527–1532, 2013. See page 4.

[68] Soyoung Youn, Tae Kim, In-Young Yoon, Jahyun Jeong, Hye Young Kim, Ji Won Han, Jong-Min Kim, and Ki Woong Kim. Progression of cognitive impairments in idiopathic rem sleep behaviour disorder. *Journal of Neurology, Neurosurgery & Psychiatry*, 87(8):890–896, 2016. See page 4.

[69] Daphné Génier Marchand, Jacques Montplaisir, Ronald B Postuma, Shady Rahayel, and Jean-François Gagnon. Detecting the cognitive prodrome of dementia with lewy bodies: a prospective study of rem sleep behavior disorder. *Sleep*, 40(1):zsw014, 2017. See page 4.

[70] Wolfgang H Jost. Gastrointestinal motility problems in patients with parkinson's disease. *Drugs & aging*, 10(4):249–258, 1997. See page 5.

[71] C Coates and AMO Bakheit. Dysphagia in parkinson's disease. *European neurology*, 38(1):49–52, 1997. See page 5.

[72] Marty Hinz, Alvin Stein, and Ted Cole. Parkinson's disease: carbidopa, nausea, and dyskinesia. *Clinical pharmacology: advances and applications*, 6:189, 2014. See page 5.

[73] Zaid S Heetun and Eamonn MM Quigley. Gastroparesis and parkinson's disease: a systematic review. *Parkinsonism & related disorders*, 18(5):433–440, 2012. See page 5.

[74] Alfonso Fasano, Naomi P Visanji, Louis WC Liu, Antony E Lang, and Ronald F Pfeiffer. Gastrointestinal dysfunction in parkinson's disease. *The Lancet Neurology*, 14(6):625–639, 2015. See page 5.

[75] Fabrizio Stocchi and Margherita Torti. Constipation in parkinson's disease. *International review of neurobiology*, 134:811–826, 2017. See page 5.

[76] LL Edwards, RF Pfeiffer, EMM Quigley, Ruth Hofman, and Mary Balluff. Gastrointestinal symptoms in parkinson's disease. *Movement disorders: official journal of the Movement Disorder Society*, 6(2):151–156, 1991. See page 5.

[77] Ehsan Adeli, Feng Shi, Le An, Chong-Yaw Wee, Guorong Wu, Tao Wang, and Dinggang Shen. Joint feature-sample selection and robust diagnosis of parkinson's disease from mri data. *NeuroImage*, 141:206–219, 2016. See pages 5, 9.

[78] Murray W Johns. A new method for measuring daytime sleepiness: the epworth sleepiness scale. *sleep*, 14(6):540–545, 1991. See page 5.

[79] Jerome A Yesavage, Terence L Brink, Terence L Rose, Owen Lum, Virginia Huang, Michael Adey, and Von Otto Leirer. Development and validation of a geriatric depression screening scale: a preliminary report. *Journal of psychiatric research*, 17(1):37–49, 1982. See page 5.

[80] Martine Visser, Johan Marinus, Anne M Stiggelbout, and Jacobus J Van Hilten. Assessment of autonomic dysfunction in parkinson's disease: the scopa-aut. *Movement disorders: official journal of the Movement Disorder Society*, 19(11):1306–1312, 2004. See page 5.

[81] Richard L Doty, Paul Shaman, Charles P Kimmelman, and Michael S Dann. University of pennsylvania smell identification test: a rapid quantitative olfactory function test for the clinic. *The Laryngoscope*, 94(2):176–178, 1984. See page 6.

[82] Aaron Smith. *Symbol digit modalities test*. Western Psychological Services Los Angeles, 1973. See page 6.

[83] Arthur L Benton, Nils R Varney, and Kerry deS Hamsher. Visuospatial judgment: A clinical test. *Archives of neurology*, 35(6):364–367, 1978. See page 6.

[84] Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699, 2005. See page 6.

[85] Ralph HB Benedict, David Schretlen, Lowell Groninger, and Jason Brandt. Hopkins verbal learning test–revised: Normative data and analysis of inter-form and test-retest reliability. *The Clinical Neuropsychologist*, 12(1):43–55, 1998. See page 6.

[86] Ralph M Reitan. Validity of the trail making test as an indicator of organic brain damage. *Perceptual and motor skills*, 8(3):271–276, 1958. See page 7.

[87] Christopher G Goetz, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, Cristina Sampaio, Glenn T Stebbins, Matthew B Stern, Barbara C Tilley, Richard Dodel, Bruno Dubois, et al. Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): process, format, and clinimetric testing plan. *Movement disorders*, 22(1):41–47, 2007. See page 7.

[88] Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Progress in neurobiology*, 95(4):629–635, 2011. See page 8.

[89] Kenneth Marek, Sohini Chowdhury, Andrew Siderowf, Shirley Lasch, Christopher S Coffey, Chelsea Caspell-Garcia, Tanya Simuni, Danna Jennings, Caroline M Tanner, John Q Trojanowski, et al. The parkinson's progression markers initiative (ppmi)–establishing a pd biomarker cohort. *Annals of clinical and translational neurology*, 5(12):1460–1477, 2018. See page 8.

[90] Susel Góngora Alonso, Isabel de la Torre-Díez, Sofiane Hamrioui, Miguel López-Coronado, Diego Calvo Barreno, Lola Morón Nozaleda, and Manuel Franco. Data mining algorithms

and techniques in mental health: A systematic review. *Journal of medical systems*, 42(9):1–15, 2018. See page 9.

[91] W Gulbinat. What is the role of who as an intergovernmental organisation. *The coordination of telematics in healthcare. World Health Organisation Geneva, Switzerland*, 1997. See page 9.

[92] David J Hand and Niall M Adams. Data mining. *Wiley StatsRef: Statistics Reference Online*, pages 1–7, 2014. See page 9.

[93] Mehrbakhsh Nilashi, Othman Ibrahim, Hossein Ahmadi, Leila Shahmoradi, and Mohammadreza Farahmand. A hybrid intelligent system for the prediction of parkinson's disease progression using machine learning techniques. *Biocybernetics and Biomedical Engineering*, 38(1):1–15, 2018. See pages 9, 11.

[94] Markus Wenzel, Fausto Milletari, Julia Krüger, Catharina Lange, Michael Schenk, Ivayla Apostolova, Susanne Klutmann, Marcus Ehrenburg, and Ralph Buchert. Automatic classification of dopamine transporter spect: deep convolutional neural networks can be trained to be robust with respect to variable image characteristics. *European journal of nuclear medicine and molecular imaging*, 46(13):2800–2811, 2019. See page 9.

[95] Alberto Llera, Ismael Huertas, Pablo Mir, and Christian F Beckmann. Quantitative intensity harmonization of dopamine transporter spect images using gamma mixture models. *Molecular imaging and biology*, 21(2):339–347, 2019. See page 9.

[96] Gurpreet Singh, Meet Vadera, Lakshminarayanan Samavedham, and Erle Chuen-Hian Lim. Machine learning-based framework for multi-class diagnosis of neurodegenerative diseases: A study on parkinson's disease. *IFAC-PapersOnLine*, 49(7):990–995, 2016. See pages 10, 11, and 46.

[97] Si-Chun Gu, Jie Zhou, Can-Xing Yuan, and Qing Ye. Personalized prediction of depression in patients with newly diagnosed parkinson's disease: A prospective cohort study. *Journal of affective disorders*, 268:118–126, 2020. See page 10.

[98] Xi Zhang, Jingyuan Chou, Jian Liang, Cao Xiao, Yize Zhao, Harini Sarva, Claire Henchcliffe, and Fei Wang. Data-driven subtyping of parkinson's disease using longitudinal clinical records: a cohort study. *Scientific reports*, 9(1):1–12, 2019. See page 10.

[99] Filippo Cavallo, Alessandra Moschetti, Dario Esposito, Carlo Maremmani, and Erika Rovini. Upper limb motor pre-clinical assessment in parkinson's disease using machine learning. *Parkinsonism & related disorders*, 63:111–116, 2019. See page 10.

[100] Enas Abdulhay, N Arunkumar, Kumaravelu Narasimhan, Elamaran Vellaiappan, and V Venkatraman. Gait and tremor investigation using machine learning techniques for the diagnosis of parkinson disease. *Future Generation Computer Systems*, 83:366–373, 2018. See page 10.

[101] C Kotsavasiloglou, N Kostikis, Dimitrios Hristu-Varsakelis, and M Arnaoutoglou. Machine learning-based classification of simple drawing movements in parkinson's disease. *Biomedical Signal Processing and Control*, 31:174–180, 2017. See page 11.

[102] Deepak Gupta, Shirsh Sundaram, Ashish Khanna, Aboul Ella Hassanien, and Victor Hugo C De Albuquerque. Improved diagnosis of parkinson's disease using optimized crow search algorithm. *Computers & Electrical Engineering*, 68:412–424, 2018. See page 11.

[103] Andreas Kuhner, Isabella Katharina Wiesmeier, Massimo Cenciarini, Timo Leon Maier, Stefan Kammermeier, Volker Arnd Coenen, Wolfram Burgard, and Christoph Maurer. Motion biomarkers showing maximum contrast between healthy subjects and parkinson's disease patients treated with deep brain stimulation of the subthalamic nucleus. a pilot study. *Frontiers in neuroscience*, 13:1450, 2020. See page 11.

[104] Jefferson S Almeida, Pedro P Rebouças Filho, Tiago Carneiro, Wei Wei, Robertas Damaševičius, Rytis Maskeliūnas, and Victor Hugo C de Albuquerque. Detecting parkinson's disease with sustained phonation and speech signals using machine learning techniques. *Pattern Recognition Letters*, 125:55–62, 2019. See page 11.

[105] Kemal Polat. Classification of parkinson's disease using feature weighting method on the basis of fuzzy c-means clustering. *International Journal of Systems Science*, 43(4):597–609, 2012. See page 11.

[106] Yang Wang, An-Na Wang, Qing Ai, and Hai-Jing Sun. An adaptive kernel-based weighted extreme learning machine approach for effective detection of parkinson's disease. *Biomedical Signal Processing and Control*, 38:400–410, 2017. See page 11.

[107] Orhan Yaman, Fatih Ertam, and Turker Tuncer. Automated parkinson's disease recognition based on statistical pooling method using acoustic features. *Medical hypotheses*, 135:109483, 2020. See page 11.

[108] Lizbeth Naranjo, Carlos J Perez, Jacinto Martin, and Yolanda Campos-Roca. A two-stage variable selection and classification approach for parkinson's disease detection by using voice recording replications. *Comput. Methods Programs Biomed.*, 142:147–156, 2017. See page 11.

[109] Marek Wodzinski, Andrzej Skalski, Daria Hemmerling, Juan Rafael Orozco-Arroyave, and Elmar Nöth. Deep learning approach to parkinson's disease detection using voice recordings and convolutional neural network dedicated to image classification. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 717–720. IEEE, 2019. See page 11.

[110] Resul Das. A comparison of multiple classification methods for diagnosis of parkinson disease. *Expert Systems with Applications*, 37(2):1568–1572, 2010. See page 11.

[111] Andoni Angulo Celada. *Predicción del diagnóstico y evolución de la enfermedad de Parkinson mediante características no motoras.* Master thesis, 2020. See page 15.

[112] Harry Khamis. Measures of association: how to choose? *Journal of Diagnostic Medical Sonography*, 24(3):155–162, 2008. See page 18.

[113] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003. See page 21.

[114] Yiming Yang and Jan O Pedersen. A comparative study on feature selection in text categorization. In *Icml*, volume 97, page 35. Nashville, TN, USA, 1997. See page 21.

[115] Ryan J Urbanowicz, Melissa Meeker, William La Cava, Randal S Olson, and Jason H Moore. Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, 85:189–203, 2018. See page 21.

[116] Yishi Zhang, Shujuan Li, Teng Wang, and Zigang Zhang. Divergence-based feature selection for separate classes. *Neurocomputing*, 101:32–42, 2013. See page 21.

[117] Mark Andrew Hall. Correlation-based feature selection for machine learning. 1999. See page 22.

[118] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014. See page 22.

[119] Tom M Mitchell et al. Machine learning. 1997. See page 22.

[120] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992. See page 23.

[121] Shan Suthaharan. Support vector machine. In *Machine learning models and algorithms for big data classification*, pages 207–235. Springer, 2016. See page 23.

[122] J Ross Quinlan. *C4. 5: programs for machine learning.* Elsevier, 2014. See page 23.

[123] Jesús M Pérez, Javier Muguerza, Olatz Arbelaitz, Ibai Gurrutxaga, and José I Martín. Combining multiple class distribution modified subsamples in a single tree. *Pattern Recognition Letters*, 28(4):414–422, 2007. See page 23.

[124] Pat Langley, Wayne Iba, Kevin Thompson, et al. An analysis of bayesian classifiers. In *Aaai*, volume 90, pages 223–228. Citeseer, 1992. See page 24.

[125] James Joyce. Bayes' theorem. 2003. See page 24.

[126] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009. See page 24.

[127] Sagar Sharma. Activation functions in neural networks. *towards data science*, 6, 2017. See page 24.

[128] Henry Leung and Simon Haykin. The complex backpropagation algorithm. *IEEE Transactions on signal processing*, 39(9):2101–2104, 1991. See page 24.

[129] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989. See page 24.

[130] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. See page 24.

[131] Min Xu, Pakorn Watanachaturaporn, Pramod K Varshney, and Manoj K Arora. Decision tree regression for soft classification of remote sensing data. *Remote Sensing of Environment*, 97(3):322–336, 2005. See page 24.

[132] Steven Abney. Bootstrapping. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 360–367, 2002. See page 24.

[133] William W Cohen. Learning trees and rules with set-valued features. In *AAAI/IAAI, Vol. 1*, pages 709–716, 1996. See page 24.

[134] Swarnalatha Purushotham and BK Tripathy. Evaluation of classifier models using stratified tenfold cross validation techniques. In *International Conference on Computing and Communication Systems*, pages 680–690. Springer, 2011. See page 24.

[135] Claude Sammut and Geoffrey I Webb. *Encyclopedia of machine learning*. Springer Science & Business Media, 2011. See pages 24, 25.

[136] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013. See page 26.

[137] John Kruschke. Doing bayesian data analysis: A tutorial with r, jags, and stan. 2014. See page 26.

[138] Giorgio Corani and Alessio Benavoli. A bayesian approach for comparing cross-validated algorithms on multiple data sets. *Machine Learning*, 100(2):285–304, 2015. See page 26.

[139] Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *The Journal of Machine Learning Research*, 18(1):2653–2688, 2017. See page 26.

[140] John K Kruschke and Torrin M Liddell. The bayesian new statistics: Two historical trends converge. *SSRN Electronic Journal*, 2606016, 2015. See page 26.

[141] Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 856–863, 2003. See page 28.

[142] Ali Haghpanah Jahromi and Mohammad Taheri. A non-parametric mixture of gaussian naive bayes classifiers based on local independent features. In *2017 Artificial Intelligence and Signal Processing Conference (AISP)*, pages 209–212. IEEE, 2017. See page 30.

[143] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016. See page 30.

[144] Nahathai Wongpakaran, Tinakon Wongpakaran, and Pimolpun Kuntawong. Evaluating hierarchical items of the geriatric depression scale through factor analysis and item response theory. *Heliyon*, 5(8):e02300, 2019. See page 60.

Appendix A

This appendix explains the variables used in the supervised database, as well as their semantic grouping.

In total, four versions of the database are created, in which the variables are grouped (or ungrouped) semantically, depending on the test or questionnaire to which they belong. In addition, there are also four variables containing general and clinical patient information. Table 1 summarises the variables used.

Gender

The Gender attribute has only one variable: GENDER. This variable can take the following values: 0 (woman with reproductive capacity), 1 (woman without reproductive capacity) and 2 (man).

Years of education

The variable that indicates the age of education is EDUCYRS. This can have values from 5 years to 26, always using discrete numbers.

Dominant hand

Only one variable is used to identify the dominant hand: HANDED. This variable can take a value of 1 if it is right-handed, a value of 2 if it is left-handed and a value of 3 if it is ambidextrous.

Age

A variable with the same name is used to indicate age. This variable has a range of 54.09, starting at 31.2 and ending at 85.29. Therefore, the criterion in PPMI of being older than 30 years is satisfied.

Class

The attribute class is the variable we want to classify in the supervised analysis. This variable matches with the attribute name and with the possible values. Therefore, the values that this target variable can have are: IDIOPATHIC PD (subjects with Idiopathic Parkinson's disease) and HC (healthy control subjects).

Epworth Sleepiness Scale (ESS)

The variables of the ESS questionnaire are the answers to each question. In total there are 8 questions that are recorded from variable ESS1 to variable ESS8, with the variable number corresponding to the question. The values of the variables range from 0 to 3, these being the subject's answers. In the second grouping (Intermediate 2), all responses are added together to create the variable ESS_total.

Geriatric Depression Scale (GDS)

The variables used in the GDS questionnaire are the scores of the answers from the 15 questions asked to the subjects. Therefore, the value is 1 if the response denotes a symptom of depression and 0 otherwise. The variable name always starts with “GDS" and ends with a word indicating the topic of the question. When the variables are grouped, the article by Wongpakaran *et al.* [144] was taken into account, where it mentions that 6 of the 15 questions are enough to predict a person’s depression. For this reason, the first grouping is done with these questions and the variable GDS_6 has been created. To create this variable, the scores of the responses have been summed. Finally, in the last version, all response scores are added up, i.e. how many depressive symptoms the subject has.

SCOPA-AUT

As happens with the other two questionnaires, there is one variable for each SCOPA_AUT question. The name of these variables always starts with “SCAU" and is followed by the question number. The variables “SCAUSEX1" and “SCAUSEX2" are the answers of the questions about sex, i.e. the first variable corresponds to questions 22 and 24, while the second variable corresponds to the combination of questions 23 and 25. These questions are asked according to gender, men are asked questions 22 and 23, while women are asked questions 24 and 25. The possible values for these variables are the following: 0 (“never"), 1 (“sometimes"), 2 (“regularly"), 3 (“often") and 4 (“use catheter" or “not applicable"). Questions SCAU23A, SCAU26A, SCAU26B and SCAU26C are about medicines. These take a value of 0 if the patient does not take the respective medicine and a value of 1 otherwise. In the case of SCAU23A it asks only for men, so the value for women is always 0. Finally, the variable PTCBOTH indicates who has filled in the test and takes the following values: 1 (Patient), 2 (Caregiver) and 3 (Patient and Caregiver).

In the first grouping, the values are added up depending on the domain and the following variables are created: SCAU_gastroint, SCAU_urinary, SCAU_cardiovascular, SCAU_thermoreg and SCAU_sexual. In the final version, all variables are added together. The variable PTCBOTH is only used in the individual version, because the medical experts indicated so.

Symbol Digit Modalities Test (SDMT)

The SDMT test has a single variable: SDMTOTAL. The value of this variable corresponds to the number of symbols that have been correctly identified.

Benton Judgment of Line Orientation Test (BJLOT)

In the BJLOT test, one variable is used for each item. These variables have two values: 1 if the two lines with the fan have been identified correctly and 0 otherwise. The variable names have the prefix “BJLOTPAR" followed by the item number. They have been grouped in the Intermediate 2, adding the values of the items and the BJLOT_total variable has been created.

University of Pennsylvania Smell Identification Test (UPSIT)

The UPSIT test contains 4 variables, one for each booklet. These variables take the value of the score (how many smells have been identified), so it can have values from 0 to 10. The variable name is formed by “UPSITBK" + the number of the booklet. In the second grouping, a variable called UPSIT_total is created by adding up the scores of all the booklets.

Montreal Cognitive Assessment (MoCA)

In the MoCA test, variables are created for each possible point, with a total of 27 variables. All variable names start with “MCA" and are followed by a code that identifies the item and what is being tested in that item. For example, “MCACLCKN" is one of the variables of the clock-drawing task (“CLCK") and indicates whether they have correctly drawn the numbers (“N") on the clock. MoCA_total is the sum of all items. If the patient has less than 12 years of education, an extra point is added. The maximum possible is 30 points.

First, the variables are grouped by the cognitive domain that each section assesses, adding up the scores, and the following variables are created: moca_visuo, moca_naming, moca_attention, moca_verbal, moca_recall and moca_orientation.

The language part is divided into two tests: the first consists of repeating two sentences said by the examiner and the score obtained is stored in the variable MCASNTNC; the second test consists of saying the highest number of words beginning with ‘P’ during one minute. This second test creates two variables: MCAVFNUM indicates how many words the subject has said, while MCAVF scores 0 if the subject says less than 11 words or 1 otherwise. The latter (MCAVF) will be used to calculate moca_verbal and MCATOT, but will not be included in the individual version of the database, to avoid redundancy with MCAVFNUM.

Hopkins Verbal Learning Test - Revised (HVLTR)

The variables used for the HVLTR test correspond to the different assessments:

- The variables HVLTRT1, HVLTRT2 and HVLTRT3 are the amount of correctly recalled words in the three learning trials. They have a value between 0 (no words correct) and 12 (all words remembered).
- The variable HVLTRDLY contains the sum of the remembered words in the retrieval test. The values are the same as above.
- The variable HVLTREC corresponds to the words recognized in the recognition test. It has values from 0 to 12.

A single grouping is made by adding together the correct words in the learning trials (HVLTRT_total). From a clinical point of view, it has been considered appropriate to leave the HVLTRDLY and HVLTREC variables separate, as these items on the same neuropsychological test represent a cognitive ability that does not overlap with HVLTRT_total.

Total (15)	**Inter. 2 (19)**	**Inter. 1 (53)**	**Individual (107)**
GENDER	GENDER	GENDER	GENDER
EDUCYRS	EDUCYRS	EDUCYRS	EDUCYRS
HANDED	HANDED	HANDED	HANDED
AGE	AGE	AGE	AGE
ESS_total	ESS_total	ESS1..8	ESS1..8
GDS_total	GDS_6	GDS_6	GDSSATIS, GDSBORED, GDSGSPIR, GDSHLPLS, GDSWRTLS, GDSHOPLS
	–	–	GDSDROPD, GDSEMPTY, GDSAFRAD, GDSHAPPY, GDSHOME, GDSMEMRY, GDSALIVE, GDSENRGY, GDSBETER
SCAU_total	SCAU_gastroint	SCAU_gastroint	SCAU1..7
	SCAU_urinary	SCAU_urinary	SCAU8..13
	SCAU_cardiovascular	SCAU_cardiovascular	SCAU14..16
	SCAU_thermoreg	SCAU_thermoreg	SCAU17..18, SCAU20..21
	SCAU19	SCAU19	SCAU19
	SCAU_sexual	SCAU_sexual	SCAUSEX1..2, SCAU23A
	SCAU26A	SCAU26A	SCAU26A
	SCAU26B	SCAU26B	SCAU26B
	SCAU26C	SCAU26C	SCAU26C
–	–	–	PTCGBOTH
SDMTOTAL	SDMTOTAL	SDMTOTAL	SDMTOTAL
BJLOT_total	BJLOT_total	BJLOTPAR1..15	BJLOTPAR1..15
UPSIT_total	UPSIT_total	UPSITBK1..4	UPSITBK1..4
MoCA_total	moca_visuo	moca_visuo	MCAALTTM, MCACUBE, MCACLCKC, MCACLCKN, MCACLCKH
	moca_naming	moca_naming	MCALION, MCARHINO, MCACAMEL
	moca_attention	moca_attention	MCAFDS, MCABDS, MCAVIGIL, MCASER7
	moca_verbal	moca_verbal	MCASNTNC, MCAVF
	MCAABSTR	MCAABSTR	MCAABSTR
	moca_recall	moca_recall	MCAREC1..5
	moca_orientation	moca_orientation	MCADATE, MCAMONTH, MCAYR, MCADAY, MCAPLACE, MCACITY
–	–	–	MCAVFNUM
HVLTRT_total	HVLTRT_total	HVLTRT_total	HVLTRT1..3
HVLTRDLY	HVLTRDLY	HVLTRDLY	HVLTRDLY
HVLTREC	HVLTREC	HVLTREC	HVLTREC
Class	Class	Class	Class

Table 1: Summary of the variables used in each version of the supervised databases.

Appendix B

This appendix presents the descriptions and graphs of the variables. In the case of qualitative variables, i.e., CLASS, GENDER and HANDED variables, they are represented by pie charts, while quantitative variables are represented by box plots and histograms.

Class

The distribution of the Class can be seen in Figure 1, represented by a char pie. In blue colour are the HC subjects, in total there are 197 subjects. The pink part belongs to the IDIOPATHIC PD subjects, exactly 490 patients.

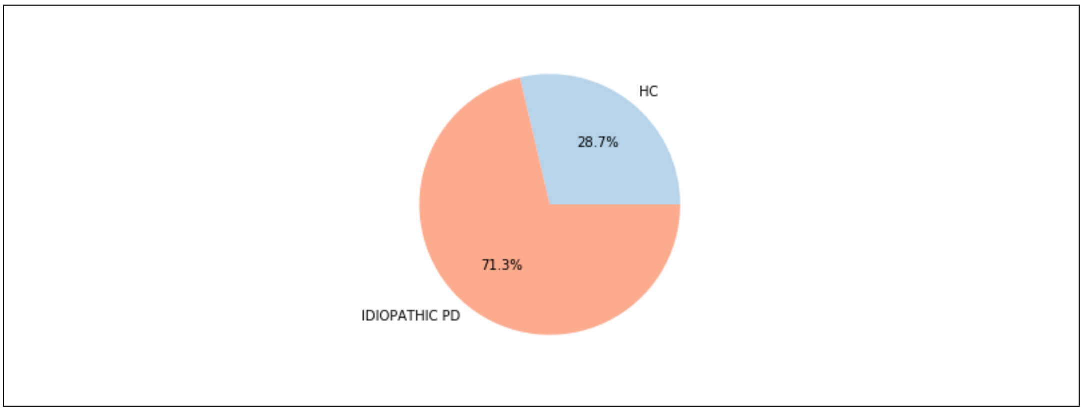


Figure 1: Class pie chart

Gender

In Figure 2 we can see the distribution of the GENDER variable. The majority of the subjects are men, exactly 448 subjects. It is followed by women without reproductive capacity, 201 subjects. Finally, there are 38 subjects who are women with reproductive capacity. In all categories the class distribution is similar, where the majority of subjects are IDIOPATHIC PD.

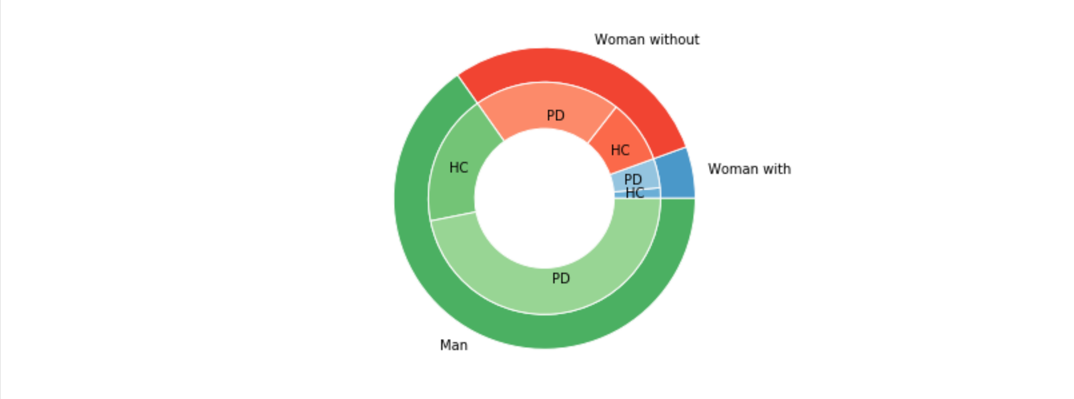
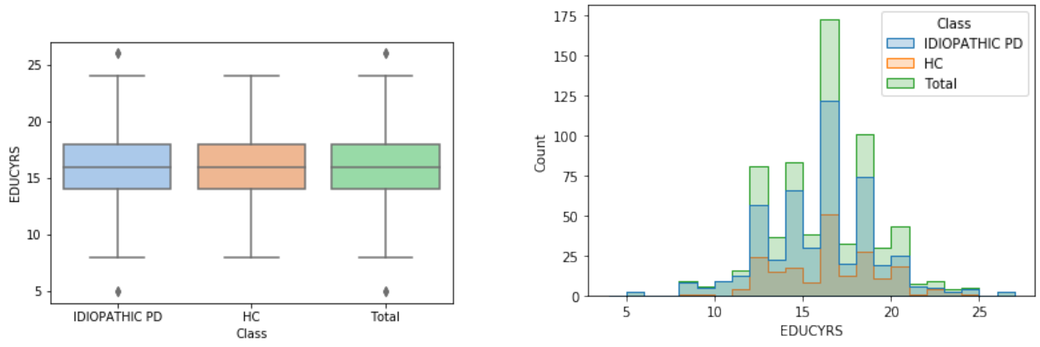


Figure 2: GENDER pie chart

Years of education

In Figure 3, the EDUCYRS variable is represented by a box plot and a histogram. In the Total box plot (the one that includes IDIOPATHIC PD and HC subjects), it can be seen that most of the subjects have studied between 14 and 18 years, with a median of 16 years. The lowest age is 5 years and the highest age is 26 years, but these two subjects are outliers. Looking at the distribution of the classes (the HC and IDIOPATHIC PD charts), the ages are evenly distributed, without taking into account outliers.



(1) EDUCYRS box plot

(2) EDUCYRS histogram

Figure 3: Graphics of Years of education variable

Dominant hand

A pie chart has been used to represent the variable HANDED, as can be seen in Figure 4. The majority of the subjects are right-handed (591 subjects), followed by left-handed (72 subjects) and finally ambidextrous (24 subjects). The distribution of classes is similar in each of them.

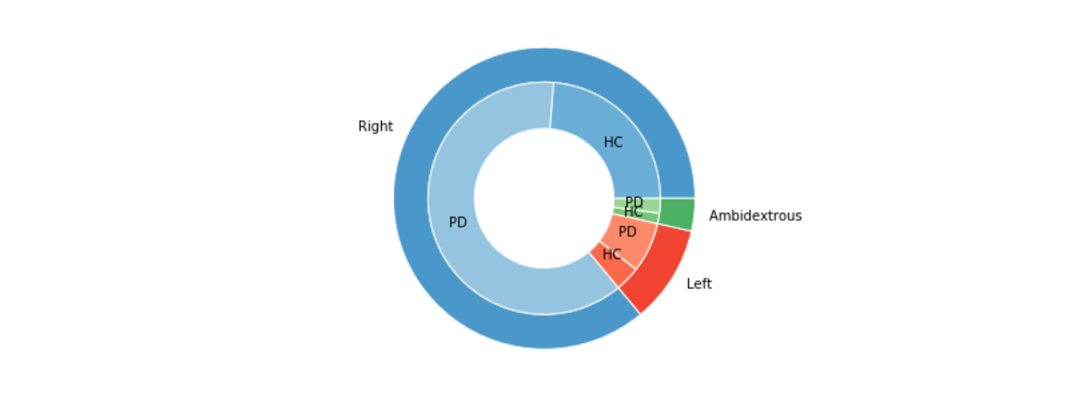
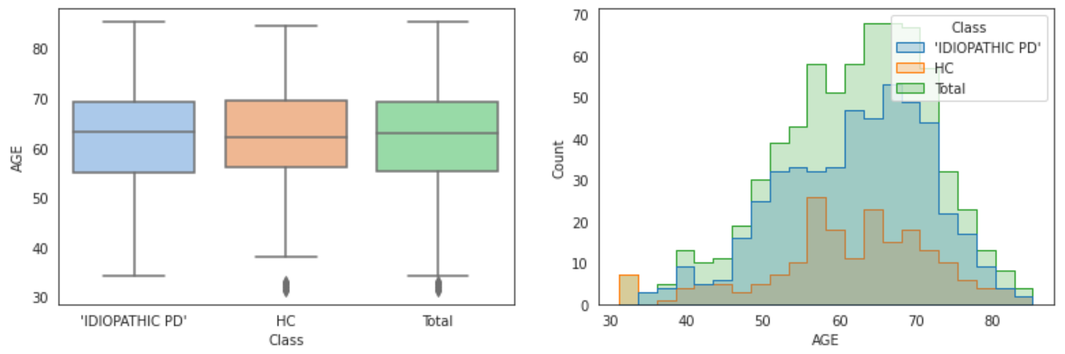


Figure 4: HANDED pie chart

Age

The graphs representing the variable AGE can be seen in Figure 5. In the Total box plot there are outliers, each with a value below 34.12. These outliers correspond to the HC subjects, as can be seen in the box plot. It can also be seen in the histogram how these data (in orange) are separated from the general group. In relation to the general distribution, both classes have a similar distribution.



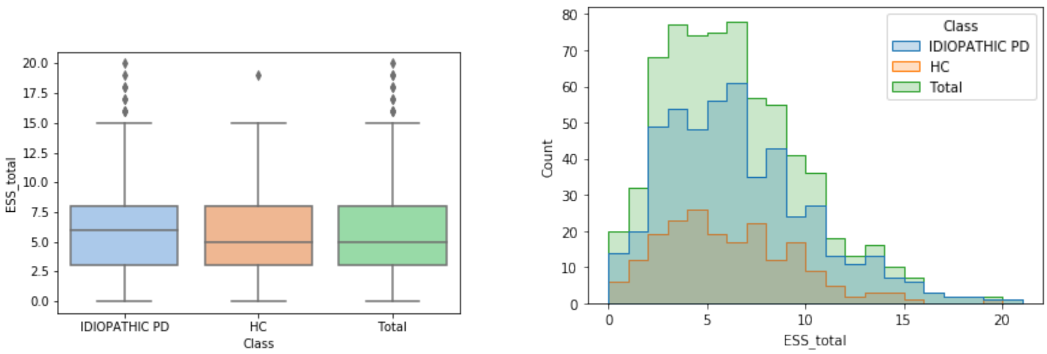
(1) AGE box plot

(2) AGE histogram

Figure 5: Graphics of age variable

Epworth Sleepiness Scale (ESS)

In Figure 6 we can see the box plot and the histogram corresponding to the ESS_total (the sum of all the questions). On the one hand, in the first graph, the interquartile range of both classes is the same, although the median of IDIOPATHIC PD is higher. On the other hand, there are more outliers that are IDIOPATHIC PD. Therefore, there are more subjects who tend to fall asleep during daily activities while having Parkinson's disease. This difference can also be seen in the histogram, with the blue tail being longer than the orange tail.



(1) ESS_total box plot

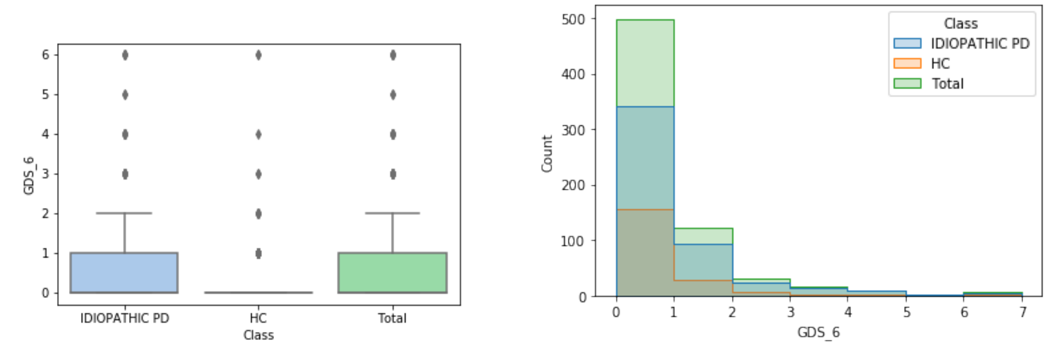
(2) ESS_total histogram

Figure 6: Graphics of Epworth Sleepiness Scale variable

Geriatric Depression Scale (GDS)

The variables of the Geriatric Depression Scale are different in the Intermediate 2 and the Total versions. The first one uses the variable GDS_6 and the second one uses the variable GDS_total. These variables indicate how many depressive symptoms the patient has.

Figure 7 shows the representation of the variable. Most of the HC subjects present no symptoms of depression (GDS_6 = 0), although there are 5 atypical patients. As for the IDIOPATHIC PD subjects, although most of these have no symptoms, it is normal for them to have 1 or 2, although above this value they are also considered outliers.

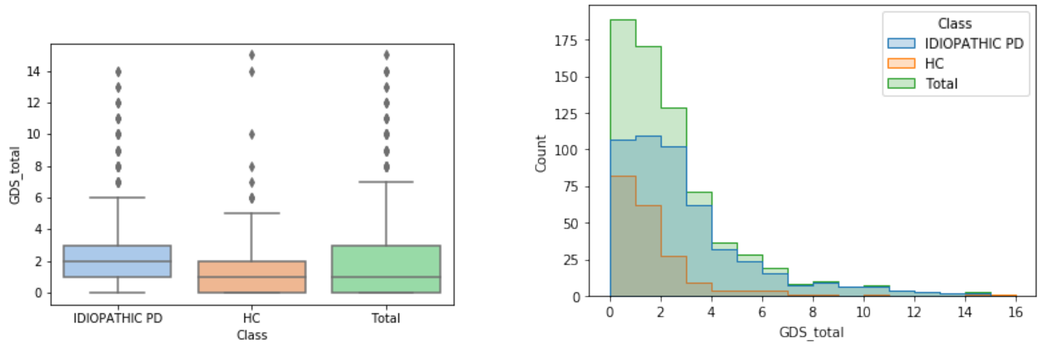


(1) GDS_6 box plot

(2) GDS_6 histogram

Figure 7: Graphics of Geriatric Depression Scale variable in Intermediate 2

Figure 8 is the representation of GDS_total. Compared to the previous one, the medians are higher: for IDIOPATHIC PD it is 2 and for HC it is 1. Comparing the outliers, in this case we have more than in the previous case. Although, for both GDS_6 and GDS_total, it seems that HC subjects have fewer symptoms than IDIOPATHIC PD subjects, as we can see in the histograms.



(1) GDS_total box plot

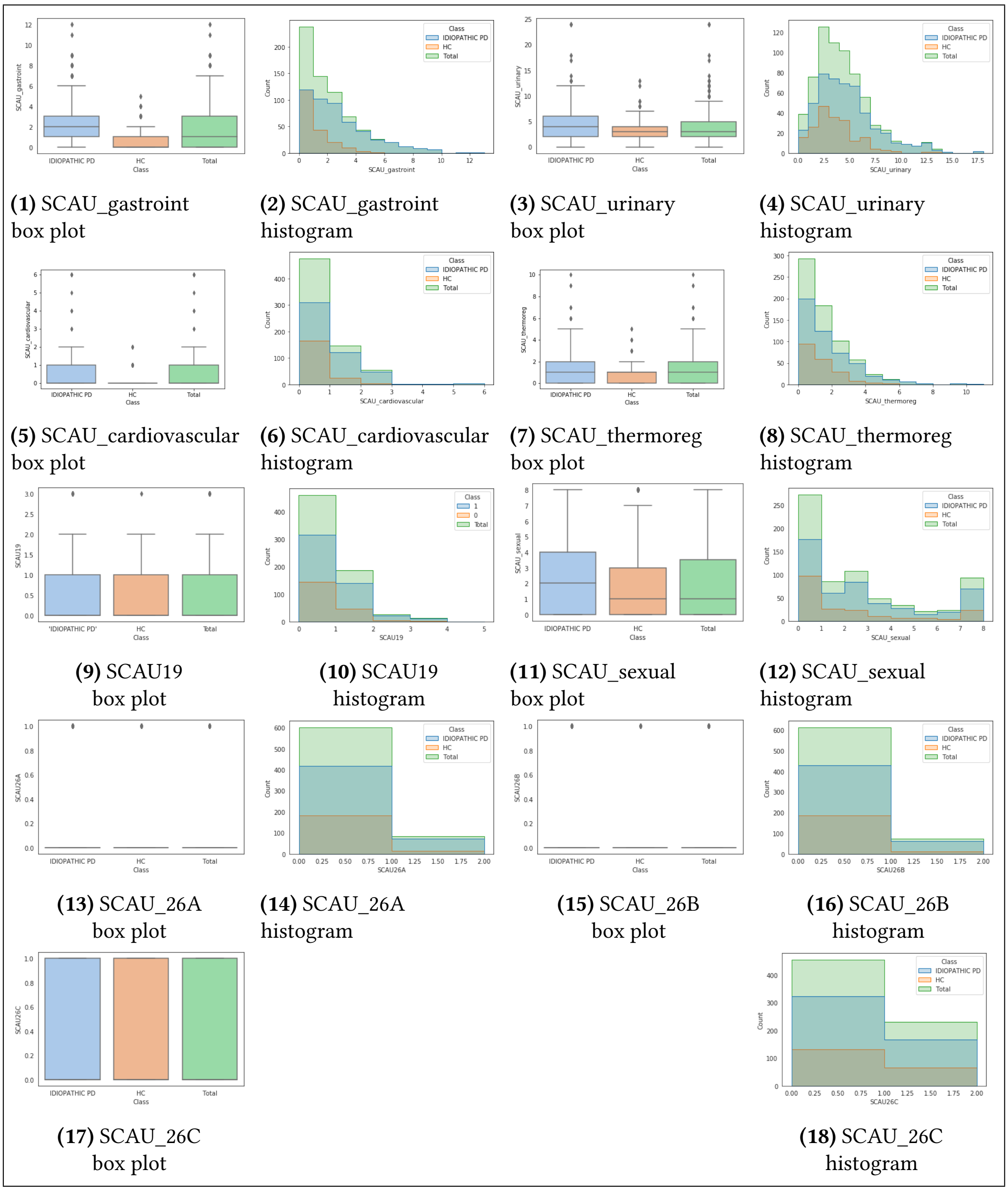
(2) GDS_total histogram

Figure 8: Graphics of Geriatric Depression Scale variable in Total

SCOPA-AUT

The SCOPA-AUT questionnaire is first grouped by sections and then a general grouping is made. The first grouping is found in Intermediate 2 (Figure 9) and the final grouping in the total version (Figure 10).

In most sections it can be observed that HC has a lower median than IDIOPATHIC PD (in the case of the box plots) or the right tail of the histogram is longer. For the case of SCAU_cardiovascular, the median is the same, but the interquartile range is greater in IDIOPATHIC PD than in HC. There are special cases (SCAU19, SCAU_26A, SCAU26_B, SCAU26_C) where there is no difference between classes.



(1) SCAU_gastroint box plot

(2) SCAU_gastroint histogram

(3) SCAU_urinary box plot

(4) SCAU_urinary histogram

(5) SCAU_cardiovascular box plot

(6) SCAU_cardiovascular histogram

(7) SCAU_thermoreg box plot

(8) SCAU_thermoreg histogram

(9) SCAU19 box plot

(10) SCAU19 histogram

(11) SCAU_sexual box plot

(12) SCAU_sexual histogram

(13) SCAU_26A box plot

(14) SCAU_26A histogram

(15) SCAU_26B box plot

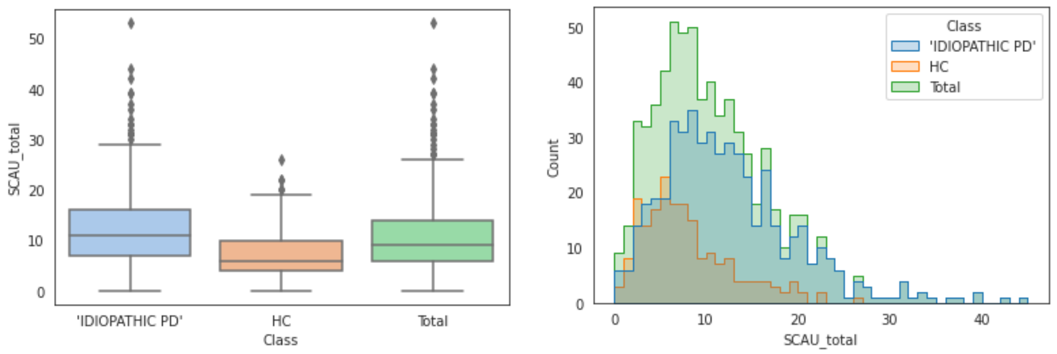
(16) SCAU_26B histogram

(17) SCAU_26C box plot

(18) SCAU_26C histogram

Figure 9: Graphics of SCOPA-AUT variables in Intermediate 2

Figure 10 gives us a more general idea of the questionnaire. As mentioned before, HC has a lower median than IDIOPATHIC PD. If we look at the outliers, there is a clear clustering between 5 and 12 (if we look at the Total boxplot), but from 22.5 above are outliers.



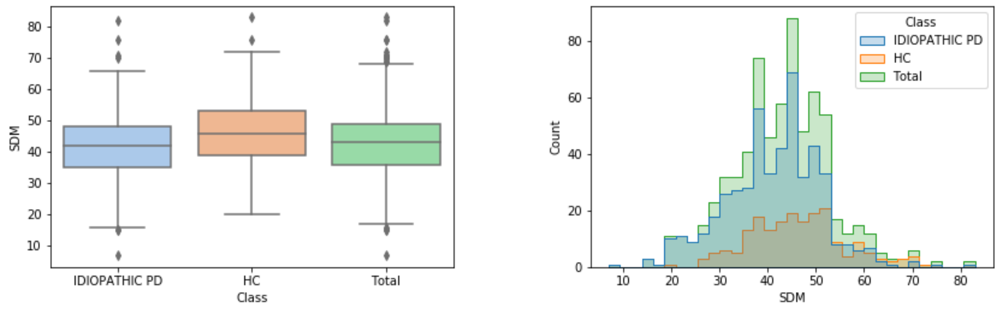
(1) SCAU_total box plot

(2) SCAU_total histogram

Figure 10: Graphics of SCOPA-AUT in Total

Symbol Digit Modalities Test (SDMT)

The representation of the SDMTOTAL variable can be found in Figure 11. In the box plot and in the histogram it can be seen that the HC class has a higher score than the IDIOPATHIC PD class. Therefore, in general, control patients tend to get more symbols right. The lowest score is 7 points and the highest score is 83, i.e. no participant has correctly identified all 120 symbols.



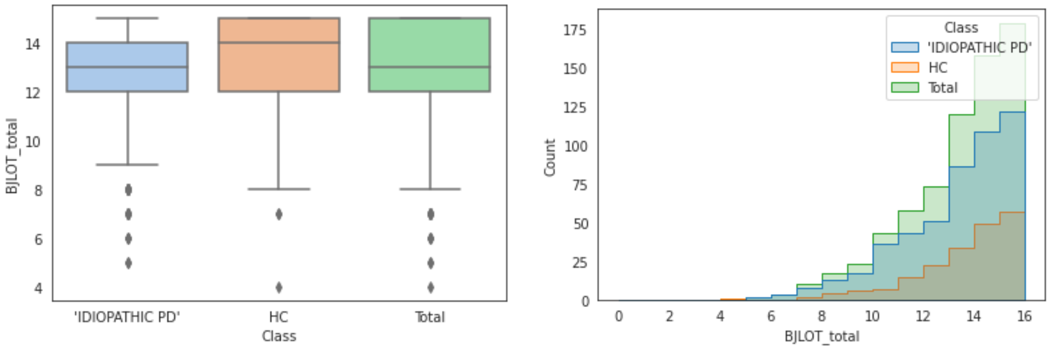
(1) SDMTOTAL box plot

(2) SDMTOTAL histogram

Figure 11: Graphics of Symbol Digit Modalities Test variable

Benton Judgment of Line Orientation Test (BJLOT)

The scores of the different subjects in the Benton Judgment of Line Orientation Test are shown in Figure 12. We can see how the subjects have a median of 13 and the average HC subjects 14. Most of the subjects were able to do more than 8 items correctly (see Total box plot), although there are 4 subjects who obtained a lower score.



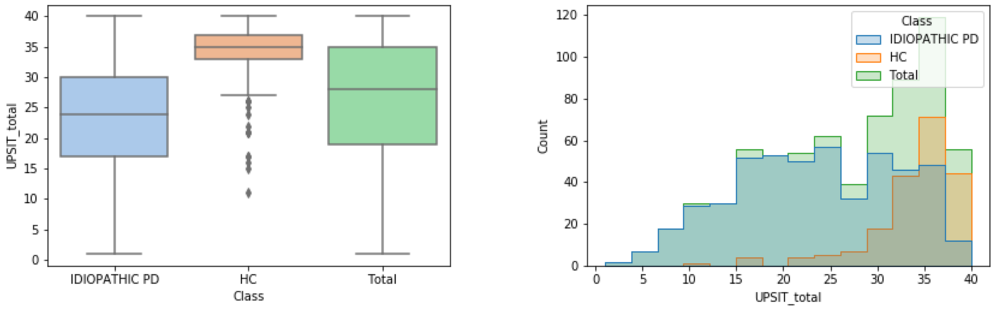
(1) BJLOT_total box plot

(2) BJLOT_total histogram

Figure 12: Graphics of Benton Judgment of Line Orientation Test variable

University of Pennsylvania Smell Identification Test (UPSIT)

The graphs corresponding to the UPSIT variable are shown in Figure 13. In the box plot we see that the IDIOPATHIC PD subjects have a median of 24, i.e. they correctly perform a little more than half of the questions. The HC subjects, however, have a median of 35. In the histogram these subjects are mostly on the right side.



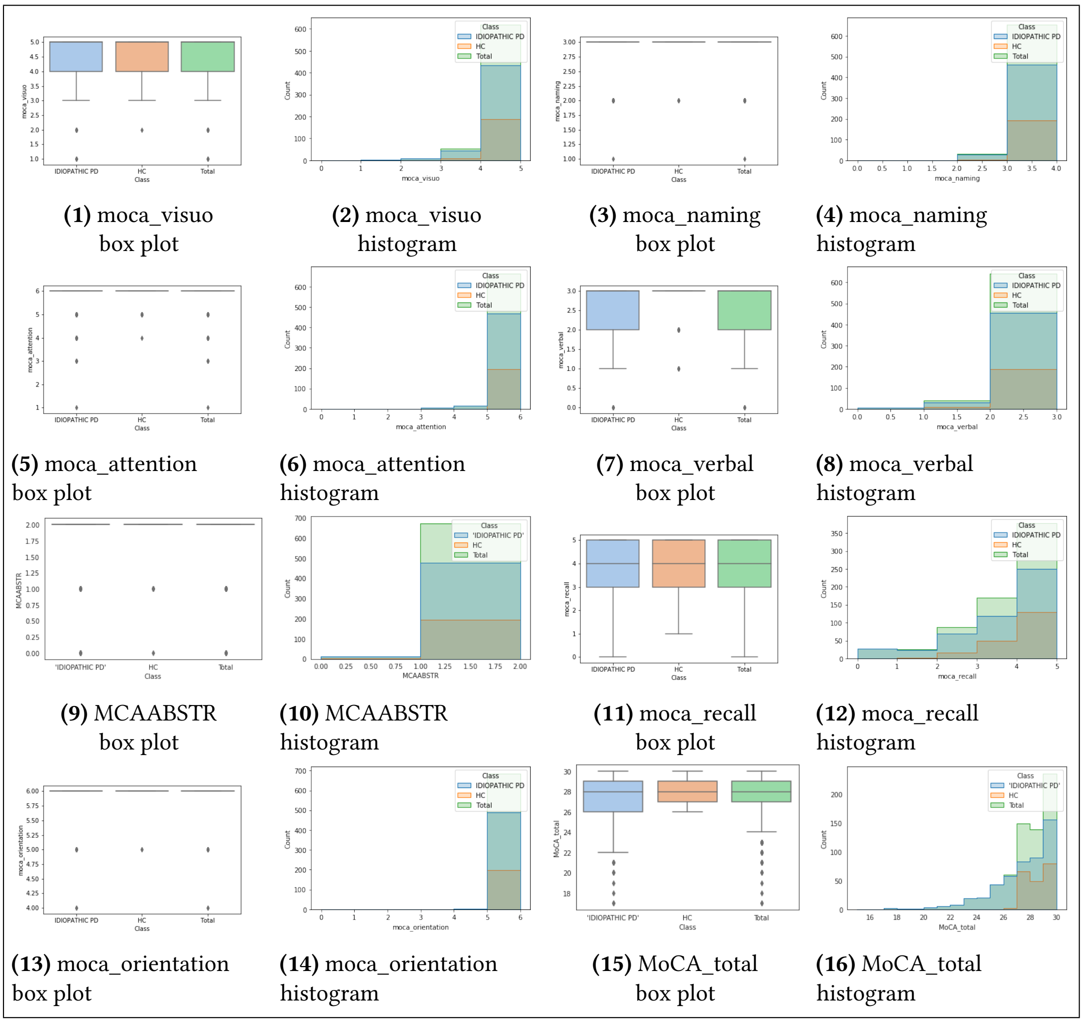
(1) UPSIT_total box plot

(2) UPSIT_total histogram

Figure 13: Graphics of University of Pennsylvania Smell Identification Test variable

Montreal Cognitive Assessment (MoCA)

The variables of the Intermediate 2 version are visualised in Figure 14. There are also the graphs of MoCA_total (variable of the total and Intermedia 2 versions), in which we can see that the variance of the HCs is smaller than that of IDIOPATHIC PD, although they have the same median. In the case of moca_verbal, the interquartile range is greater in IDIOPATHIC PD than in HC, although the median is the same. In the other graphs we have identical box plots for both classes. However, for moca_recall, there is no HC subject who scored 0.



(1) moca_visuo box plot

(2) moca_visuo histogram

(3) moca_naming box plot

(4) moca_naming histogram

(5) moca_attention box plot

(6) moca_attention histogram

(7) moca_verbal box plot

(8) moca_verbal histogram

(9) MCAABSTR box plot

(10) MCAABSTR histogram

(11) moca_recall box plot

(12) moca_recall histogram

(13) moca_orientation box plot

(14) moca_orientation histogram

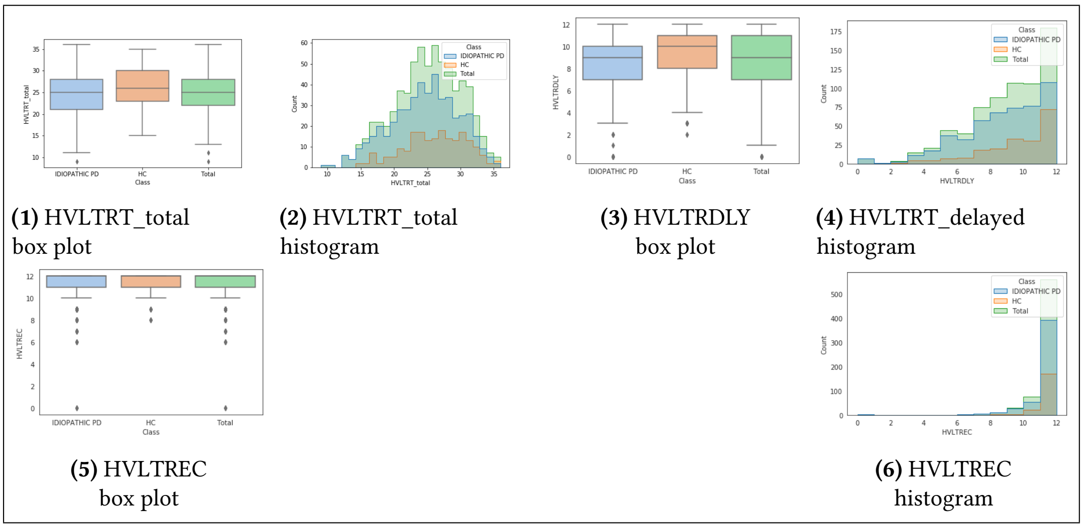
(15) MoCA_total box plot

(16) MoCA_total histogram

Figure 14: Graphics of Montreal Cognitive Assessment variables in Intermediate 2

Hopkins Verbal Learning Test - Revised (HVLTR)

Figure 15 shows the variables of the Intermediate 2 version. There is no figure for the variables of the Total version, since they are the same variables as in the previous version but with a different name. In HVLTRT_total and HVLTRDLY variables we found more correct words for the HC subjects than for the IDIOPATHIC PD subjects, obtaining a higher value in the variables. In the HVLTREC graph, the classes are only differentiated by the outliers.



(1) HVLTRT_total box plot

(2) HVLTRT_total histogram

(3) HVLTRDLY box plot

(4) HVLTRT_delayed histogram

(5) HVLTREC box plot

(6) HVLTREC histogram

Figure 15: Graphics of Hopkins Verbal Learning Test - Revised variables

Appendix C

This subsection shows the correlations between the variables using a heat map. Three tables have been created: the first one for all the subjects, the second one for HC subjects and the last one taking into account all the IDIOPATHIC PD subjects. On the other hand, correlations greater than 0.5 or less than -0.5 will be explained.

As expected, variables created due to one-hot encoding are negatively correlated. These correlations are ignored in the following description.

Intermediate 2 version

Figure 16 shows the correlations of all subjects in the Intermediate 2 version. First of all, the variables HVLTRT_total and HVLTRDLY are positively correlated, with a correlation of 0.75. The Class variable correlates mainly with the UPSIT_total variable, with a negative correlation of -0.53.

In Figure 17, only HC subjects are taken into account. Regarding HVLTRT_total with HVLTRDLY, the correlation is maintained with the same value and the correlation between HVLTRDLY and HVLTREC is increased (0.5). Finally, the correlation between AGE and SDMTOTAL becomes more noticeable, with a value of -0.51.

Finally, the correlations between the variables were evaluated using only subjects with PD, as can be seen in the graph in Figure 18. This time no new correlations stand out, the correlation between HVLTRT_total and HVLTRDLY the value decreases a little to 0.74.

Total version

The correlations between the variables in the Total version of the database can be seen in Figure 19. Apart from the correlations highlighted in Intermediate 2, no other variables are highlighted. Therefore, there is not a high correlation between the tests.

Figure 20 shows the correlation between the variables in the Total version only taking into account the HC subjects. On the other hand, the Figure 21 shows what happens with PD subjects only. In both cases, no correlation different from the one highlighted in the Intermediate 2 version was found.

Nominal variables

In Figure 22 we find the correlations between the nominal variables, calculated using Goodman and Kruskal's lambda. They have no relationship with each other or with the class.

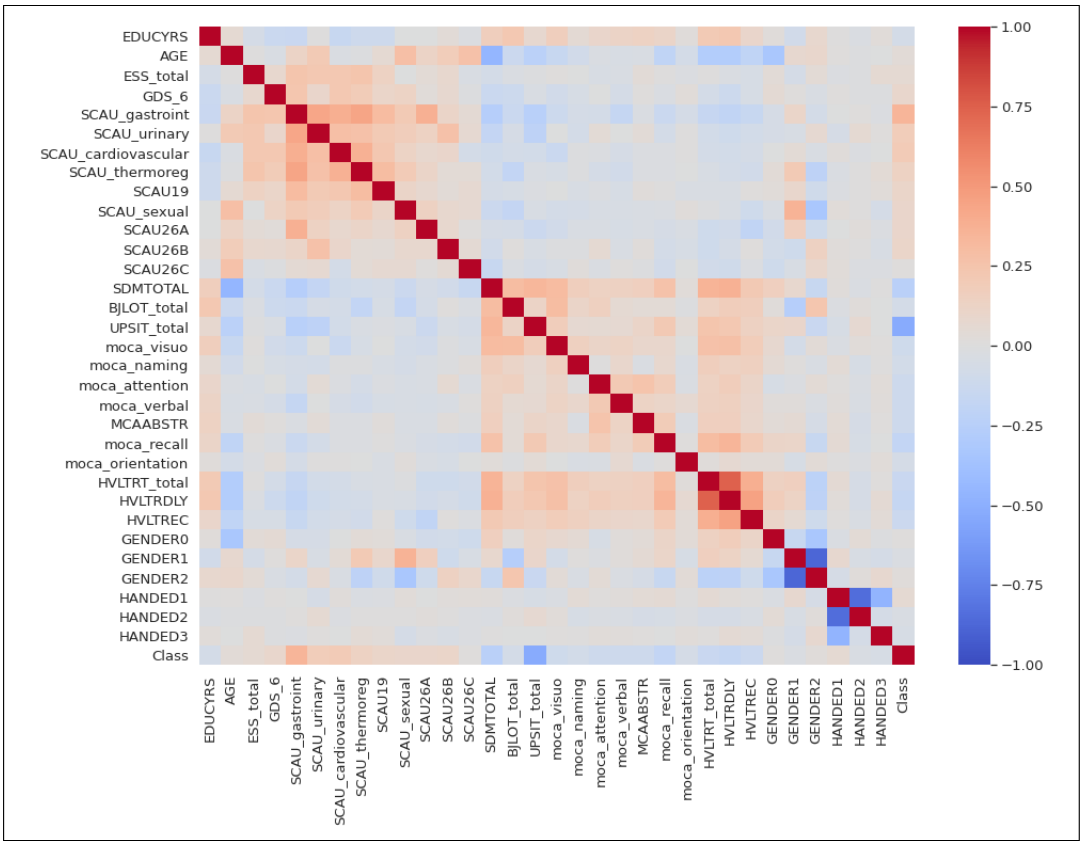


Figure 16: Heat map of correlations between the variables in the Intermediate 2 version.

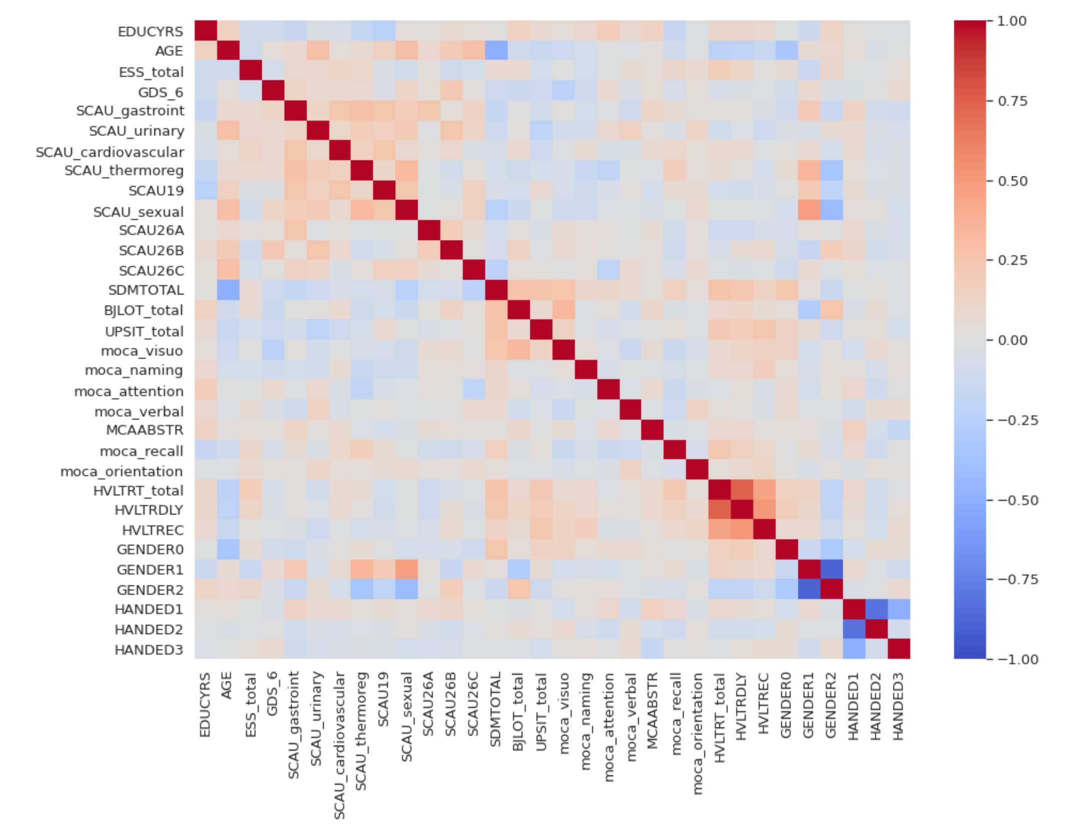


Figure 17: Heat map of correlations between variables in the Intermediate 2 version using only HC subjects.

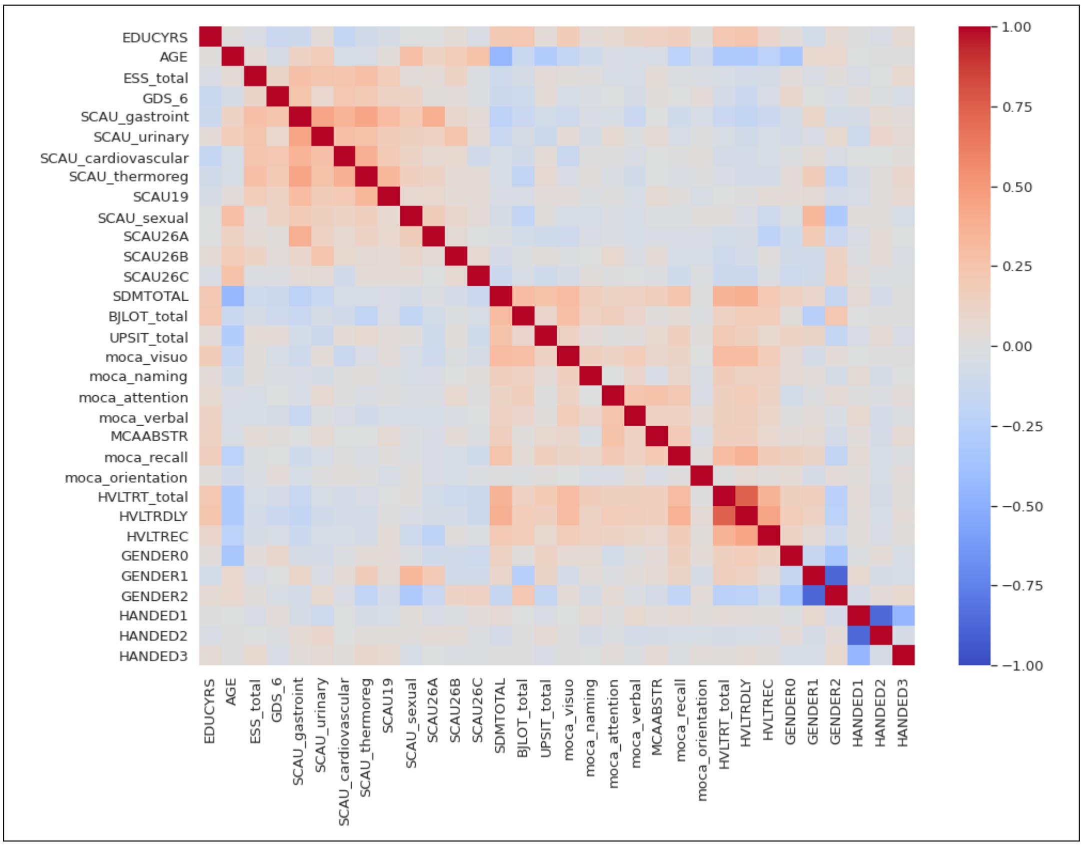


Figure 18: Heat map of correlations between variables in the Intermediate 2 version using only PD subjects.

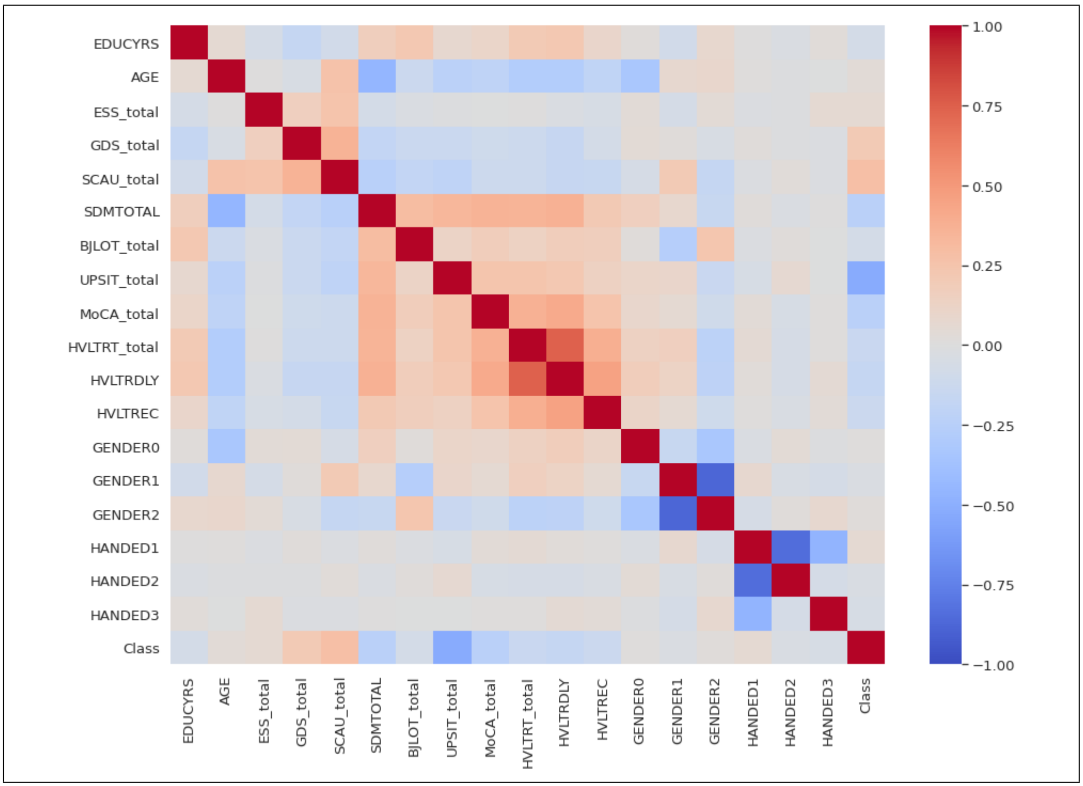


Figure 19: Heat map of correlations between the variables in the Total version.

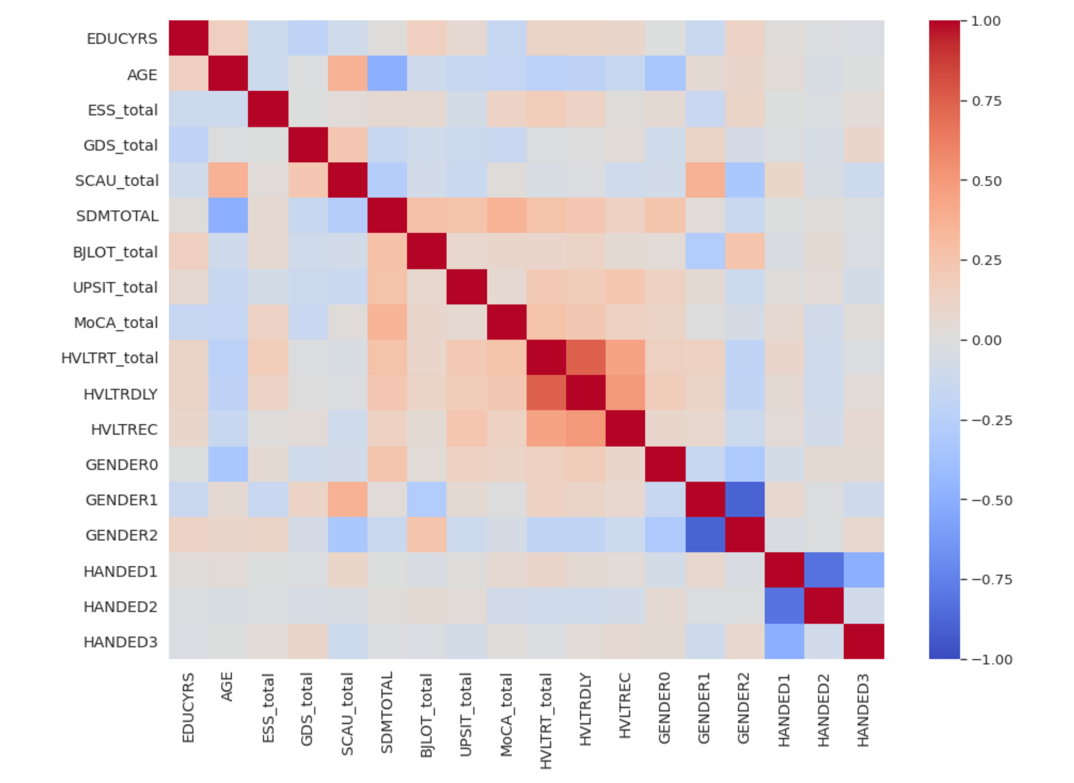


Figure 20: Heat map of correlations between variables in the Total version using only HC subjects.

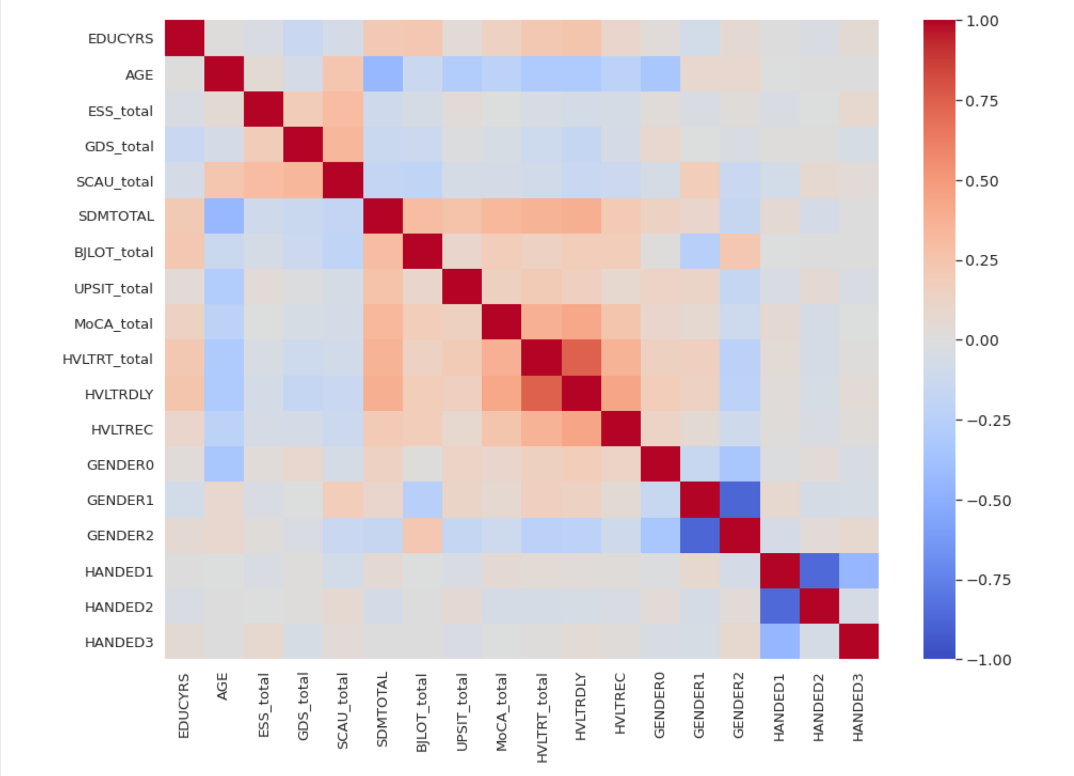
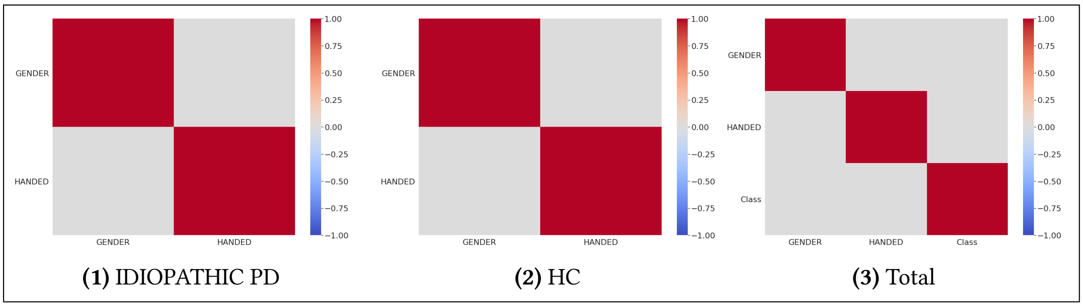


Figure 21: Heat map of correlations between variables in the Total version using only PD subjects.



(1) IDIOPATHIC PD

(2) HC

(3) Total

Figure 22: Heat map of correlations between the nominal variables.