

PARKINSON'S DISEASE ONSET DETECTION USING MACHINE LEARNING

Submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering degree in Electronics and Communication Engineering

By

Vanaparthi Sahith kumar (39130499)

Volvoji Sai Vamshi (39130515)



**DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
SCHOOL OF ELECTRICAL AND ELECTRONICS**

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC

JEPPIAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119



DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **VANAPARTHI SAHITH KUMAR** (Reg. No. 39130499) and **VOLVOJI SAI VAMSHI** (Reg.No. 39130515) who carried out the project entitled "**PARKINSON'S DISEASE ONSET DETECTION USING MACHINE LEARNING**" under our supervision from November 2022 to April 2023.

A handwritten signature in black ink, appearing to read "A. Gau", with a horizontal line underneath.

Internal Guide

Dr. A. Sahaya Anselin Nisha, M.E., Ph.D.

A handwritten signature in black ink, appearing to read "T. Ravi", with the date "15/4/2023" written below it.

Head of the Department

Dr. T. Ravi, M.E., Ph.D.

Submitted for Viva Voce Examination held on 20/04/2023

A handwritten signature in blue ink, appearing to read "Dr. Ravi", with a horizontal line underneath.

INTERNAL EXAMINER

A handwritten signature in black ink, appearing to read "Sai", with a horizontal line underneath.

EXTERNAL EXAMINER

DECLARATION

We **VANAPARTHI SAHITH KUMAR** (Reg. No.39130499) and **VOLVOJI SAI VAMSHI** (Reg. No.39130515) hereby declare that the Project Report entitled “**PARKINSON’S DISEASE ONSET DETECTION USING MACHINE LEARNING**” done by us under the guidance of **Dr.A.SAHAYA ANSELIN NISHA M.E., Ph.D.** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Electronics and Communication Engineering.

DATE : 20/04/2023

PLACE: CHENNAI

1. 
2. 

SIGNATURE OF CANDIDATE

ACKNOWLEDGEMENT

We are pleased to acknowledge our sincere thanks to the **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. We are grateful to them.

We convey our thanks to **Dr. N. M. Nandhitha, M.E., Ph.D., Dean, School of Electronics and Communication Engineering** and **Dr. T. Ravi, M.E., Ph.D., Head of the Department, Department of Electronics and Communication Engineering** for providing me necessary support and details at the right time during the progressive reviews.

We would like to express our sincere and deep sense of gratitude to our project guide **Dr. A. Sahaya Anselin Nisha, M.E., Ph.D.** for her valuable guidance, suggestions, and constant encouragement paved way for the successful completion of our project work.

We wish to express our thanks to all teaching and non-teaching staff members of the **Department of Electronics and Communication Engineering** who were helpful in many ways for the completion of the project.

We would like to thank our beloved parents for their constant love and support, without whom this would not be possible and finally we would like to thank the Almighty for His wonderful blessing.

ABSTRACT

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.

Parkinson's disease is a progressive disorder that affects the nervous system and the parts of the body controlled by the nerves. Symptoms start slowly. The first symptom may be a barely noticeable tremor in just one hand. Tremors are common, but the disorder may also cause stiffness or slowing of movement. Machine learning algorithms for the classification of PD and healthy controls or patients with comparable clinical presentations have been adopted to solve these issues and refine the diagnosis and assessment procedures of PD.

Medical observations and assessment of clinical indicators, including the identification of a variety of motor symptoms, are often used to diagnose Parkinson's disease (PD). Traditional diagnostic procedures, on the other hand, may be vulnerable to subjectivity because they rely on the assessment of motions that are sometimes subtle to human sight and hence difficult to define, potentially leading to misdiagnosis. Meanwhile, early non-motor symptoms of Parkinson's disease can be minor and be caused by a variety of other illnesses. As a result, these symptoms are frequently missed, making early PD diagnosis difficult. Machine learning algorithms for the classification of PD and healthy controls or patients with comparable clinical presentations have been adopted to solve these issues and refine the diagnosis and assessment procedures of PD.

The goals of this study are to provide a comprehensive overview of all published studies that applied machine learning models to the diagnosis of PD for an exhaustive overview of data sources, data types.

CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	iv
	LIST OF FIGURES	viii
1	INTRODUCTION	1
	1.1 INTRODUCTION	1
	1.2 BRIEF HISTORY	2
	1.3 PROBLEM STATEMENT	4
	1.3.1 AIM OF PROJECT	5
2	LITERATURE SURVEY	6
	2.1 SUMMARY	8
3	INVESTIGATION	9
	3.1 EXISTING SYSTEM	9
	3.2 PROPOSED SYSTEM	9
4	REQUIRED COMPONENTS	12
	4.1 SOFTWARE REQUIREMENTS	12
	4.1.1 PYTHON LIBRARIES IN ML	13
	4.1.2 IMPLEMENT PYTHON IN ML	13
	4.2 HARDWARE REQUIREMENTS	14
5	MACHINE LEARNING	15
	5.1 MACHINE LEARNING ALGORITHMS	18
	5.2 CLASSIFICATION MODELS	18
	5.2.1 LOGISTIC REGRESSION CLASSIFIER	18
	5.2.2 DECISION TREE CLASSIFIER	19
	5.2.3 GAUSSIAN NAÏVE BAYES CLASSIFIER	19
	5.2.4 RANDOM FOREST CLASSIFIER	20
	5.2.5 CHOOSE ML ALGORITHMS	20
	5.2.6 CHALLENGES OF ML	20

	5.3 APPLICATIONS OF MACHINE LEARNING	21
	5.3.1 AUGMENTATION	21
	5.3.2 AUTOMATION	21
	5.3.3 FINANCE INDUSTRY	21
	5.3.4 GOVERNMENT ORGANIZATION	21
	5.3.5 HEALTHCARE INDUSTRY	21
	5.3.6 MARKETING	21
	5.4 DATA MINING	22
	5.4.1 SUPERVISED LEARNING	24
	5.4.2 UNSUPERVISED LEARNING	24
	5.4.3 ANOMALY DETECTION	27
6	MODEL BUILDING AND IMPLEMENTATION	25
	6.1 DATA COLLECTION	28
	6.1.1 DATA LOCATION	29
	6.2 DATA REPRESENTATION	29
	6.3 DATA NORMALISATION	34
	6.4 CHOOSING MODEL	36
	6.5 TRAINING THE MODEL	39
	6.6 MODEL EVALUTION	40
	6.6.1 PREDICTIVE SYSTEM	42
	6.7 RESULT	42
7	CONCLUSION	44
	7.1 CONCLUSION	44
	7.2 APPLICATIONS	44
	7.3 ADVANTAGES	44
	7.4 DISADVANTAGES	44

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE NO.
1.1	MODEL OVERVIEW	5
5.1	TRADITIONAL PROGRAMMING	16
5.2	MACHINE LEARNING	16
5.3	LEARNING PHASE OF ML	17
5.4	INFERENCE MODEL OF ML	17
5.5	MACHINE LEARNING ALGORITHMS	18
6.1	UPLOADING DATASET	31
6.2	READING DATASET	31
6.3	DESCRIPTION OF DATASET	32
6.4	ATTRIBUTES AVERAGE	32
6.5	ATTRIBUTE VALUES	33
6.6	ATTRIBUTE VALUES	33
6.7	DATA NORMALISATION	36
6.8	FLOW CHART	36
6.9	MODEL TRAINING	39
6.10	MODEL EVALUTION	41
6.11	PREDICTIVE SYSTEM	42
6.12	OUTPUT	43

CHAPTER 1

INTRODUCTION

1.1 Introduction

Parkinson's disease is a neurodegenerative central nervous system syndrome that affects movement and causes tremors and stiffness. It comprises five stages and affects more than 1 million people in India each year. This is a long-term condition for which there is currently no cure. It's a neurodegenerative disease that affects the brain's dopamine-producing neurons. Parkinson's disease (PD) is a common neurological disease that affects roughly 1% of the global population over age of 55 and above. Parkinson's disease affects an estimated five million people globally. PD by 2030, the prevalence is predicted to double. Two or more cardinal motor symptoms, such as rest tremor, bradykinesia (slow movement), or rigidity (stiffness), are used to identify Parkinson's disease (stiffness). Bradykinesia is a common symptom of Parkinson's disease that responds well to current treatments. Bradykinesia is a symptom of a movement illness such as Parkinson's disease. It's caused by low dopamine levels in the brain, and it's usually detected first by family and friends. Reduced mobility is a symptom of Parkinson's disease rather than a symptom caused by it.

Symptoms occur gradually, and may begin with a barely detectable tremor in only one hand. Tremors are common, however they are often accompanied by stiffness or slowed mobility. The patient's ability to move and speak is severely limited by this disease. The patient develops a habit of falling frequently, resulting in numerous traumas. As a result, it's critical to keep track of the disease's severity and warn either the patients or their career.

Alzheimer's disease, Parkinson's disease, Arthritis disease, Dementia with Lewy bodies, Cortico basal degeneration, Progressive supranuclear palsy, and Prion disorders are among the neurodegenerative illnesses now identified. Parkinson's disease is the second most frequent disease after Alzheimer's among all of these neurodegenerative and coordinating bodily movement diseases.

Prior to cognitive and behavioural problems, movement-related symptoms such as tremor, rigidity, and difficulties in starting might be seen. PD has a significant impact on patients' quality of life (QoL), social functions, and family relationships, as well as imposing significant financial costs on individuals and society.

Traditionally, motor symptoms have been used to diagnose Parkinson's disease. Despite the fact that cardinal indications of Parkinson's disease have been established in clinical assessments, most of the rating scales used to determine disease severity have not been well reviewed and validated. Non-motor symptoms (such as cognitive and behavioural abnormalities, sleep difficulties, and sensory abnormalities such as olfactory dysfunction) are present in many individuals prior to the beginning of PD, but they lack specificity, are difficult to measure, and/or vary from patient to patient. As a result, nonmotor symptoms cannot currently be used to diagnose Parkinson's disease on their own, albeit some have been utilised as diagnostic criteria.

Parkinson's disease is a neurological condition that affects the brain. It causes tremors in the body and hands, as well as stiffness in the body. At this moment, there is no proper cure or treatment available. Only when the condition is caught early, or at its inception, can it be treated. Not only will this lower the cost of the condition, but it may also save a life.

1.2 Brief History

Machine learning has risen to prominence in the healthcare industry in recent years. Machine learning, as the name suggests, allows a computer programme to learn and extract meaningful representations from data with little or no human intervention. Machine learning models have been used to diagnose Parkinson's disease using a variety of data modalities, such as movement data 3 (such as handwriting or gait), neuroimaging, speech, cerebrospinal fluid (CSF), cardiac scintigraphy, serum, and optical coherence tomography (OCT). In the diagnosis of Parkinson's disease, machine learning also allows for the combination of diverse modalities, such as magnetic resonance imaging (MRI) and single-photon emission computed tomography (SPECT) data.

It can therefore be find significant traits that are not normally employed in the clinical diagnosis of PD using machine learning algorithms, and rely on these alternative measures to diagnose PD in preclinical stages or atypical forms. The number of papers on the use of machine learning in the diagnosis of Parkinson's disease has risen in recent years. Although prior research examined the application of machine learning in the diagnosis and assessment of Parkinson's disease, they were confined to the examination of motor symptoms, kinematics, and wearable sensor data, or studies published between 2015 and 2016.

The goals of this study are to provide a comprehensive overview of all published studies that applied machine learning models to the diagnosis of PD for an exhaustive overview of data sources, data types, machine learning models, and associated outcomes, assess and compare the feasibility and efficiency of different machine learning methods in the diagnosis of PD, and provide machine learning practitioners interested in the diagnosis of PD with an overview of previously used PD machine learning methods.

As a result, applying machine learning to clinical and non-clinical data from various modalities has frequently resulted in high diagnostic accuracies in human participants, which may encourage the use of machine learning algorithms and novel biomarkers in clinical settings to help with more accurate and informed decision making.

1.2.1 Parkinson's disease:

The majority of available methods can detect Parkinson's disease in its advanced stages, which indicates a loss of approximately 60% dopamine in the basal ganglia, which is responsible for coordinating bodily movement with a tiny amount of dopamine. In the United Kingdom, over 145,000 people have been identified alone suffering, while in India, nearly one million people are affected by the sickness, which is rapidly spreading throughout the world.

Other symptoms that a person with Parkinson's disease may experience include:

1. Depression
2. Nervousness

3. Sleeping problems and memory problems
4. Balance issues and a loss of sense of smell.

Although the exact etiology of Parkinson's disease is unknown, experts have shown that various factors contribute to the disease's onset. It consists of:

1. Genes- Researchers have discovered some extremely unusual mutant genes. Gene mutations frequently enhance the risk of Parkinson's disease, but each genetic marker has a smaller influence.
2. The environment- Certain dangerous poisons or chemical chemicals found in the environment can cause sickness, but they have a minor impact.

1.3 Problem Statement

Parkinson's disease (PD) is one of the most common neurodegenerative illnesses, affecting 1 to 2 people per 1000. It has a prevalence rate of 1% in the population over 60 years old.

Between 1990 and 2016, the estimated global population affected by PD more than quadrupled (from 2.5 million to 6.1 million), owing to an increase in the number of old persons and age-standardized prevalence rates. Parkinson's disease (PD) is a degenerative neurological ailment with both motor and non-motor symptoms that affects all elements of movement, including planning, initiation, and execution. To develop a model for predicting Parkinson's Disease based on the support vector machine, while considering sociodemographic factors, health habits, Parkinson's symptoms, sleep behavior disorders, and neuropsychiatric indicators as predictors and provide baseline data for identifying Parkinson's Disease.

1.3.1 AIM OF THE PROJECT

The main objective of this article is to understand what is Parkinson's disease and to detect the early onset of the disease. That will use here Logistic Regression, Decision Tree, Gaussian Naïve Bayes Classifier, Random Forest Algorithm and utilize the data-set. It is intended to develop more classification models, such as logistic regression, Naive Bayes, Decision Trees, and others, as well as fine-tune the model's parameters. These models would be trained using a data set that would be carefully built when the feature engineering was completed. The present review aims at describing machine learning algorithms as they have been variably applied to different aspects of Parkinson's disease diagnosis and characterization.

Parkinson disease is one of the most serious diseases. Hence diagnosing it at an earlier stage could help prevent or reduce the effects. The machine learning classification algorithms are used to predict if a person has Parkinson disease or not, comparing different machine learning algorithm such as logistic regression, decision tree, k-nearest neighbor as well as some "Ensemble" learning techniques where we attempt to improve the accuracy by combining several models. The machine learning model can be implemented to significantly improve diagnosis method of Parkinson disease.

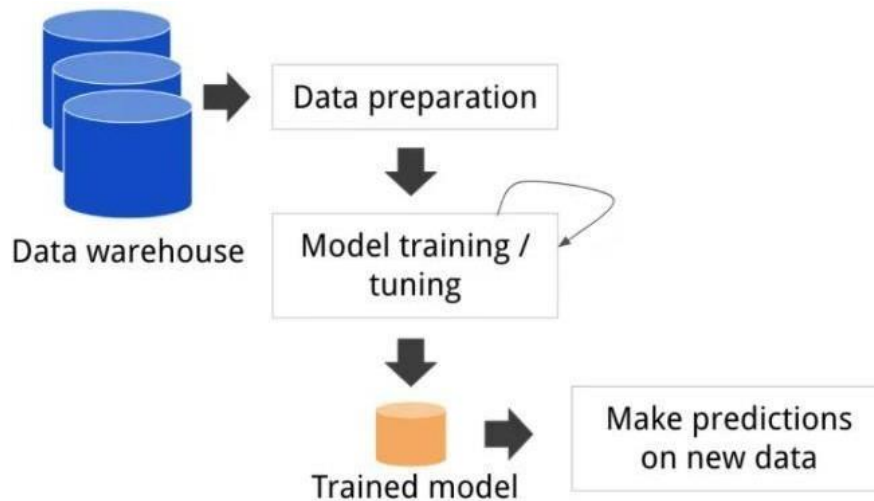


FIG No: 1.1: MODEL OVERVIEW

CHAPTER 2

LITERATURE SURVEY

Shail Raval , (2019) for the detection of PD they include all the aspects such as biological data, chemical data and genetic data [1]. In this paper they mainly focused on the symptoms like rigidity, Tremor at rest, changing voice etc. The secure data transmission is proposed through authentication check, duplication check and faulty node detection. The proposed method is applicable to long ranges of transmission. It is also supporting a retransmission concept.

Zehra Karapinar Senturk,(2017) proposed the algorithms to detect PD like support vector machine (SVM), Classification and Regression Tree (CART). It provided about 13% performance improvement for SVM, about 11% for ANN, and about 5% improvement [2]. The result shows that Naive Bayes and decision tree (j48) yield better accuracy when performed upon the discretized PD dataset with cross-validation test mode without applying any attributes selection algorithm.

Satyabrata Aich,(2018) According to this Random Forest (RF) gives more accuracy. This analysis will help the clinicians to differentiate the PD group from healthy group based on the voice data [3]. CNN's, also referred to as ConvNets, contains multiple layers and are mainly used for image processing and object detection. Yann LeCun developed the primary CNN in 1988 when it had been called LeNet. It was used for recognizing characters like ZIP codes and digits.

Timothy J. Wroge,(2017) used Extra Tree and gradient boosted Decision tree classification algorithms are used to detect variations in voice. LSTMs are a kind of Recurrent Neural Network (RNN) which will learn and memorize long-term dependencies [4]. Recalling past information for long periods is that the default behavior.

Rajalakshmi Shenbaga Moorthy used to novel analytic system for Parkinson's disease Prediction mechanism using Improved Radial Basis Function Neural Network (IRBFNN). RNNs is during a one among the deep learning models that are

used for modeling the arbitrary length sequences by applying a transition function to all or any of its hidden states during a recursive manner.[5]

Rahul R. Chakre,(2020) According to the hybrid approach, which is a combination of supervised and unsupervised techniques, is also beneficial for classification and feature extraction. Support vector machine is employed as the supervised technique for classification, and ICA is used as unsupervised technique for the feature extraction in multiclass data set.[7]. According to the field of medical diagnosis, bio-inspired computing is also a technique. Swarm intelligence and immune computing algorithms, two major subsets of bio-inspired computation, are presented for a wide range of issues. For the feature selection process, swarm intelligence approaches are described, and immune computing techniques are proposed for the classification.

Jie Mei ,(2016) used all basic algorithms of deep learning techniques for the detection of PD. Like SVM, RF, Decision Tree, ANN, KNN, Radial Basis Function Networks (RBF) and Deep Belief Networks (DBN) etc. The early identification of Parkinson's disease is critical. The identification can be performed with the use of a data mining technique. The techniques for detecting PD, such as Naive Bayes, support vector machine, multilayer perceptron neural network, and decision tree, are theoretically explained in this study. This study uses speech input from acoustic devices to predict Parkinson's disease. People from various areas and speech factors are investigated in this article in order to predict Parkinson's disease among patients. The speech dataset was used to recognize Parkinson's illness using Multi-layer Perceptron and Logistic Regression (LR) frameworks.[6]

Gabriel Solana-Lavalle,(2019) uses the algorithms such as Multilayer Perceptron (MLP), Random Forest (RF), K-Nearest Neighbor (KNN) [14]. For the prediction of Parkinson disease, three sets of experiences were conducted to obtain the features with highest contribution to PD. These three sets are

1. A population with male and female subjects (balanced),
2. Male subjects (balanced and unbalanced), and

3. Female subjects (balanced and unbalanced). In this study, the researchers used acoustic devices to collect speech parameters from 50 persons with Parkinson's disease and fifty healthy people. They employed the k-fold cross validation method for testing and claim that it can deliver 85 percent accuracy

2.1 SUMMARY

Prediction of Parkinson disorder is one of the most important problem that has to be detected in the early phases of the commencement of the disease so as to reduce the disease progression rate among the individuals .Various researches have been made to find the basic cause and some have reached to the heights by proposing a system which differentiates the healthy people from those with any ND'S (Neurodegenerative disorders) using various machine learning techniques. Lots of pre-processing feature selection and classification techniques have been implemented and developed in the past decades.

Here in our project, Machine learning algorithms for the classification of PD and healthy controls or patients with comparable clinical presentations have been adopted to solve these issues and refine the diagnosis and assessment procedures of PD

CHAPTER 3

INVESTIGATION

3.1 Existing System :

That the data must be linearly separable. The system also lacks sufficient preprocessing, visualization, and exploratory data analysis capabilities (EDA). Disadvantages of Existing System: The present systems' limitations are insufficient to deal with the complex data. In this section, we'll go through some of the flaws in the current system. The model suffers from overfitting due to a lack of data generalization, and the error on test data is large as a result of overfitting. In order to do feature engineering, the system also requires extensive data pretreatment and Exploratory Data Analysis (EDA).

3.2 Proposed System:

The present systems limitations are inadequate to deal with the complex data. In this section, we'll go through some of the issues in the current system.

- Due to a lack of data generalization, the model suffers from overfitting.
- Overfitting causes a significant level of inaccuracy on test data, and the system also necessitates extensive data pretreatment and Exploratory Data Analysis (EDA) in order to undertake feature engineering.

Medical observations and assessment of clinical indicators, including the identification of a variety of motor symptoms, are often used to diagnose Parkinson's disease (PD). Traditional diagnostic procedures, on the other hand, may be vulnerable to subjectivity because they rely on the assessment of motions that are sometimes subtle to human sight and hence difficult to define, potentially leading to misdiagnosis. Meanwhile, early non-motor symptoms of Parkinson's disease can be minor and be caused by a variety of other illnesses. As a result, these symptoms are frequently missed, making early PD diagnosis difficult. Machine learning algorithms for the classification of PD and healthy controls or patients with comparable clinical

presentations have been adopted to solve these issues and refine the diagnosis and assessment procedures of PD.

Python and Machine Learning have a connection. To create Machine Learning projects using Python, you must first understand Python and be familiar with the most extensively used Python modules.

The goal of the project was to enable researchers to better understand and assess the severity of Parkinson's disease and the characteristics of patients in order to develop more individualised treatments. The presented dataset is divided into four smaller datasets that were utilised to characterise PD characteristics in this study. Dopaminergic medicines, such as levodopa, are commonly used to reduce the symptoms of Parkinson's disease. In this project we have different features and a label to predict the Parkinson's disease. We have train and test datasets. We have features like : 'id', 'locJitter', 'locabsJitter', 'rapJitter', 'ppq5Jitter', 'ddpJitter', 'locShimmer', 'locdBShimmer', 'apq3Shimmer', 'apq5Shimmer', 'apq11Shimmer', 'ddaShimmer', 'AC', 'NTH', 'HTN', 'Median_pitch', 'Mean_pitch', 'Standard_deviation', 'Minimum_pitch', 'Maximum_pitch', 'Number_of_pulses', 'Number_of_period', 'Mean_period', 'Standard_deviation_of_period', 'Fraction_of_locally_unvoiced_frames', 'Number_of_voice_breaks', 'Degree_of_voice_breaks', 'Class', 'UPDRS'. Class is encoded with 0 and 1.

Advantages:

- The goal is to come up with new features based on the dataset's functional understanding.
- It's crucial to bear in mind that linked traits should be avoided during this process. Each feature should solely serve to improve the dataset's information.
- Visualize the dataset without cleaning it and figure out how the data is distributed.

- Identify the properties that must be handled in order to complete the data visualisation assignment.

Data is generated at an unprecedented rate as a civilization (see big data). These pieces of information can be numerical (temperature, loan amount, customer retention rate), categorised (gender, colour, highest degree achieved), or even free text (think doctor's notes or opinion polls). The practise of acquiring and analysing data from a variety of sources is known as data collection. Data must be collected and kept in a form that makes sense for the business problem at hand in order to use it to develop viable artificial intelligence (AI) and machine learning solutions. Data collection allows us to keep track of past events so that can utilise data analysis to uncover repeating patterns. Using machine learning algorithms, you may create predictive models that look for trends and forecast future changes based on those patterns

CHAPTER 4

REQUIRED COMPONENTS

4.1 Software requirements:

- Operating System: Windows 7, Windows 8, (or higher versions)
- Language: Python 3.5 and other libraries likes numpy, pandas, matplotlib, seaborn and scikit learn.
- Mozilla Firefox(or any browser)

Python in machine learning :

Despite the fact that there are several languages that may be used for machine learning, Python is, in my opinion, the greatest programming language for Machine Learning applications. This is due to the different advantages listed in the next section. R, C++, JavaScript, Java, C#, Julia, Shell, TypeScript, and Scala are some more programming languages that could be used for Machine Learning Applications. R is also an excellent language for studying machine learning.

When compared to other programming languages, Python is known for its readability and relative simplicity. Complex topics such as calculus and linear algebra are used in Machine Learning applications, which take a lot of effort and time to execute. Python helps to alleviate this effort by allowing the Machine Learning engineer to validate a concept quickly.

Python has recently established itself as the backbone of Machine Learning. In comparison to other object-oriented languages, it is a simple to learn programming language. Python is frequently used for data mining and data analysis, and it may be used to create a variety of Machine Learning models and techniques. Python is known for its readability and platform neutrality, which means you can use the same code on any machine without making any modifications. All of this makes it the ideal language for Machine Learning.

4.1.1 Python Libraries Machine Learning:

Python and Machine Learning have a connection. To create Machine Learning projects using Python, you must first understand Python and be familiar with the most extensively used Python modules. The following are the details:

SciPy:

SciPy is a Python-based programming language that includes modules for optimization, linear algebra, integration, and statistics. It is primarily used to manipulate images and do scientific calculations.

SciPy's fundamental data structure is a multi-dimensional array provided by the NumPy module. NumPy is the foundation for SciPy's array manipulation subroutines. SciPy is a Python library that provides user-friendly and powerful numerical operations while working with NumPy arrays.

NumPy:

NumPy is a Python library that can do basic numerical operations like linear algebra, Fourier transforms, and random number generation.

NumPy makes it simple to recognise arbitrary data types and integrate with most databases. NumPy can also be used as a multidimensional array for any sort of data. NumPy's strengths include a powerful N-dimensional array object, broadcasting functions, and out-of-the-box capabilities for incorporating C/C++ and Fortran code.

4.1.2 How to implement python in Machine Learning:

Python is the most widely used programming language for artificial intelligence. Python is the most widely used machine learning project on Github, and it has a long history. The majority of python programming crash courses are free to start with. Python is a straightforward programming language.

To summarise, Machine Learning is the act of identifying patterns in data and applying that information to make modifications and intelligent judgments on its own. Other

programming languages allow novices or students to learn the language before applying it to artificial intelligence or system learning.

Engineers working in machine learning and artificial intelligence prefer to create algorithms and heuristics rather than debug their code for grammatical errors. Python is an object-oriented programming language with a simple syntax and scripting capabilities. Python's scripting nature allows coders and developers to test their own theories and run calculations rapidly because it is written in a way that is virtually human-readable. That is why structural programming languages like Java, Perl, and C++, which demand complex code, aren't generally used for Machine Learning.

4.2 Hardware requirements:

- Processor: Pentium 3, Pentium 4 and higher
- RAM: 2GB/4GB RAM and higher
- Hard disk: 40GB and higher

CHAPTER 5

MACHINE LEARNING

Machine Learning is a system of computer algorithms that can learn from example through self-improvement without being explicitly coded by a programmer. Machine learning is a part of artificial Intelligence which combines data with statistical tools to predict an output which can be used to make actionable insights.

The breakthrough comes with the idea that a machine can singularly learn from the data (i.e., example) to produce accurate results. Machine learning is closely related to data mining and Bayesian predictive modeling. The machine receives data as input and uses an algorithm to formulate answers.

A typical machine learning tasks are to provide a recommendation. For those who have a Netflix account, all recommendations of movies or series are based on the user's historical data. Tech companies are using unsupervised learning to improve the user experience with personalizing recommendation.

Machine learning is also used for a variety of tasks like fraud detection, predictive maintenance, portfolio optimization, automatize task and so on.

Traditional programming differs significantly from machine learning. In traditional programming, a programmer code all the rules in consultation with an expert in the industry for which software is being developed. Each rule is based on a logical foundation; the machine will execute an output following the logical statement. When the system grows complex, more rules need to be written. It can quickly become unsustainable to maintain.



FIG:5.1 TRADITIONAL PROGRAMMING

Machine learning is supposed to overcome this issue. The machine learns how the input and output data are correlated and it writes a rule. The programmers do not need to write new rules each time there is new data. The algorithms adapt in response to new data and experiences to improve efficacy over time.



FIG:5.2 MACHINE LEARNING

Machine learning is the brain where all the learning takes place. The way the machine learns is similar to the human being. Humans learn from experience. The more we know, the more easily can predict. By analogy, when in an unknown situation, the likelihood of success is lower than the known situation. Machines are trained the same. To make an accurate prediction, the machine sees an example. When we give the machine a similar example, it can figure out the outcome. However, like a human, if its feed a previously unseen example, the machine has difficulties to predict.

The core objective of machine learning is the learning and inference. First of all, the machine learns through the discovery of patterns. This discovery is made thanks to the data. One crucial part of the data scientist is to choose carefully which data to provide to the machine. The list of attributes used to solve a problem is called a feature vector. You can think of a feature vector as a subset of data that is used to tackle a problem.

The machine uses some fancy algorithms to simplify the reality and transform this discovery into a model. Therefore, the learning stage is used to describe the data and summarize it into a model.

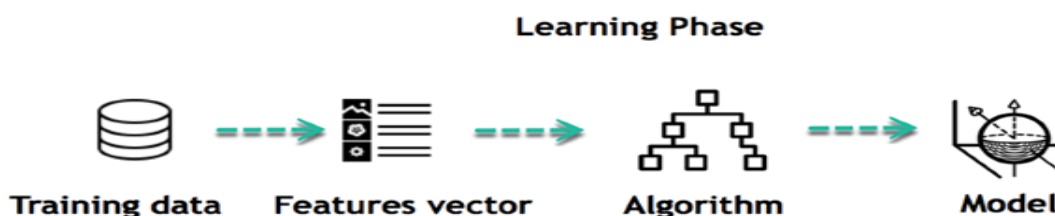


FIG: 5.3 LEARNING PHASE OF ML

For instance, the machine is trying to understand the relationship between the wage of an individual and the likelihood to go to a fancy restaurant. It turns out the machine finds a positive relationship between wage and going to a high-end restaurant.

When the model is built, it is possible to test how powerful it is on never-seen-before data. The new data are transformed into a features vector, go through the model and give a prediction. This is all the beautiful part of machine learning. There is no need to update the rules or train again the model. You can use the model previously trained to make inference on new data.

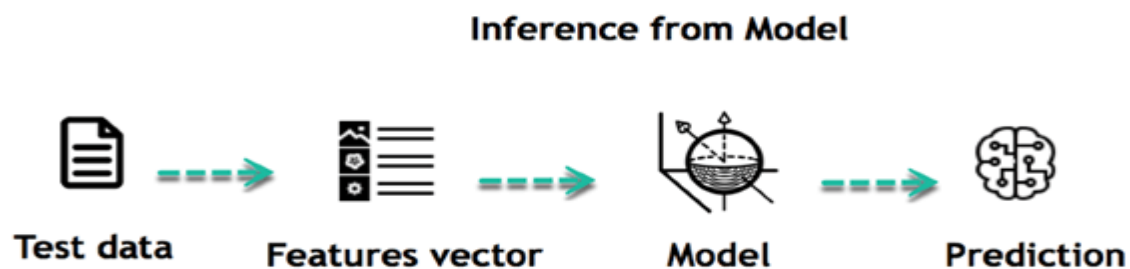


FIG: 5.4 INFERENCE MODEL OF ML

The life of Machine Learning programs is straightforward and can be summarized in the following points:

1. Define a question
2. Collect data
3. Visualize data
4. Train algorithm
5. Test the Algorithm
6. Collect feedback
7. Refine the algorithm
8. Loop 4-7 until the results are satisfying
9. Use the model to make a prediction.

5.1 Machine Learning Algorithms and Usage:

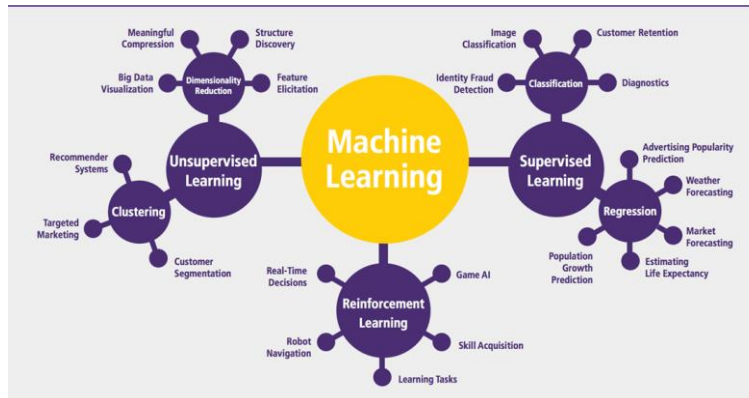


FIG: 5.5 MACHINE LEARNING ALGORITHMS

Machine learning can be grouped into two broad learning tasks: Supervised , Unsupervised. There are many other algorithms

An algorithm uses training data and feedback from humans to learn the relationship of given inputs to a given output. For instance, a practitioner can use marketing expense and weather forecast as input data to predict the sales of cans.

You can use supervised learning when the output data is known. The algorithm will predict new data.

There are two categories of supervised learning:

- Classification task
- Regression task

5.2 Classification Models:

5.2.1 Logistic Regression Classifier:

Logistic regression is a classification algorithm that uses supervised learning to predict the likelihood of a target variable. Because the nature of the target or dependent variable is dichotomous, there are only two possible classes. Simply put, the dependent variable is binary in nature, with data coded as either 1 (for success/yes) or 0 (for failure/no).

A logistic regression model predicts $P(Y=1)$ as a function of X mathematically. It is one of the most basic ML algorithms and can be used to solve a variety of classification problems such as spam detection, diabetes prediction, cancer detection, and so on.

5.2.2 Decision Tree Classifier:

Decision Trees are one type of Supervised Machine Learning (you explain what the input is and what the corresponding output is in the training data) in which the data is continuously split based on a specific parameter. The tree is explained by two entities: decision nodes and leaves. Decision trees have the following drawbacks: They are unstable, which means that a small change in the data can result in a significant change in the structure of the optimal decision tree. They are frequently erroneous. Many other predictors outperform this one with comparable data.

A Decision Tree is a supervised learning technique that may be applied to both classification and regression issues, however it is most commonly employed to solve classification problems. Internal nodes represent dataset attributes, branches represent decision rules, and each leaf node provides the conclusion in this tree-structured classifier. The Decision Node and the Leaf Node are the two nodes of a Decision tree. Leaf nodes are the output of those decisions and do not contain any more branches, whereas Decision nodes are used to make any decision and have several branches.

- The decisions or tests are based on the characteristics of the presented dataset.
- It's a graphical depiction for obtaining all feasible solutions to a problem/decision depending on certain parameters.
- It's termed a decision tree because, like a tree, it starts with a root node that grows into more branches and forms a tree-like structure.
- It utilises the CART algorithm, which stands for Classification and Regression Tree algorithm, to form a tree.
- A decision tree simply asks a question and divides the tree into subtrees based on the answer (Yes/No).

5.2.3 Gaussian Naive Bayes Classifier:

Gaussian Naive Bayes is a Naive Bayes variant that uses the Gaussian normal distribution and works with continuous data. Naive Bayes are supervised machine learning classification algorithms that are based on the Bayes theorem. It is a simple classification.

Predicting the class of the test data set is simple and quick. It also excels at multi-class prediction. When the assumption of independence is met, a Naive Bayes classifier outperforms other models such as logistic regression and requires less training data.

5.2.4 Random Forest Classifier:

Random forest is a versatile, user-friendly machine learning algorithm that produces excellent results most of the time even without hyper-parameter tuning. Because of its simplicity and versatility (it can be used for both classification and regression tasks), it is also one of the most widely used algorithms. It improves accuracy through cross validation. The random forest classifier will handle missing values while maintaining the accuracy of a large portion of the data. It will not allow over-fitting trees in the model if there are more trees.

5.2.5 How to Choose Machine Learning Algorithm

There are plenty of machine learning algorithms. The choice of the algorithm is based on the objective.

In the Machine learning example below, the task is to predict the type of flower among the three varieties. The predictions are based on the length and the width of the petal. The picture depicts the results of ten different algorithms. The picture on the top left is the dataset. The data is classified into three categories: red, light blue and dark blue. There are some groupings. For instance, from the second image, everything in the upper left belongs to the red category, in the middle part, there is a mixture of uncertainty and light blue while the bottom corresponds to the dark category. The other images show different algorithms and how they try to classify the data.

5.2.6 Challenges and Limitations of Machine Learning

The primary challenge of machine learning is the lack of data or the diversity in the dataset. A machine cannot learn if there is no data available. Besides, a dataset with a lack of diversity gives the machine a hard time.

5.3 Applications of Machine Learning:

5.3.1 Augmentation:

- Machine learning, which assists humans with their day-to-day tasks, personally or commercially without having complete control of the output. Such machine learning is used in different ways such as Virtual Assistant, Data analysis, software solutions. The primary user is to reduce errors due to human bias.

5.3.2 Automation:

- Machine learning, which works entirely autonomously in any field without the need for any human intervention. For example, robots performing the essential process steps in manufacturing plants.

5.3.3 Finance Industry

- Machine learning is growing in popularity in the finance industry. Banks are mainly using ML to find patterns inside the data but also to prevent fraud.

5.3.4 Government organization

- The government makes use of ML to manage public safety and utilities. Take the example of China with the massive face recognition. The government uses Artificial intelligence to prevent jaywalker.

5.3.5 Healthcare industry

- Healthcare was one of the first industry to use machine learning with image detection.

5.3.6 Marketing

- Broad use of AI is done in marketing thanks to abundant access to data. Before the age of mass data, researchers develop advanced mathematical tools like Bayesian analysis to estimate the value of a customer. With the boom of data, marketing department relies on AI to optimize the customer relationship and marketing campaign.

Machine learning is the best tool so far to analyze, understand and identify a pattern in the data. One of the main ideas behind machine learning is that the computer can be trained to automate tasks that would be exhaustive or impossible for a human being.

A machine can be trained to translate the knowledge of an expert into features. The features are all the characteristics of a house, neighborhood, economic environment, etc. that make the price difference. For the expert, it took him probably some years to master the art of estimate the price of a house. His expertise is getting better and better after each sale.

For the machine, it takes millions of data, (i.e., example) to master this art. At the very beginning of its learning, the machine makes a mistake, somehow like the junior salesman. Once the machine sees all the example, it got enough knowledge to make its estimation. At the same time, with incredible accuracy. The machine is also able to adjust its mistake accordingly.

Most of the big company have understood the value of machine learning and holding data. McKinsey have estimated that the value of analytics ranges from \$9.5 trillion to \$15.4 trillion while \$5 to 7 trillion can be attributed to the most advanced AI techniques.

Machine learning (ML) is the study of computer algorithms that improve automatically through experience. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

5.4 Data mining:

Machine learning and data mining often employ the same methods and overlap significantly, but while machine learning focuses on prediction, based on known properties learned from the training data, data mining focuses on the discovery of (previously) unknown properties in the data (this is the analysis step of knowledge discovery in databases). Data mining uses many machine learning methods, but with different goals; on the other hand, machine learning also employs data mining methods as "unsupervised learning" or as a preprocessing step to improve learner accuracy.

Much of the confusion between these two research communities (which do often have separate conferences and separate journals, ECML PKDD being a major exception) comes from the basic assumptions they work with: in machine learning, performance is usually evaluated with respect to the ability to reproduce known knowledge, while in knowledge discovery and data mining (KDD) the key task is the discovery of previously unknown knowledge. Evaluated with respect to known knowledge, an uninformed (unsupervised) method will easily be outperformed by other supervised methods, while in a typical KDD task, supervised methods cannot be used due to the unavailability of training data.

Machine learning also has intimate ties to optimization: many learning problems are formulated as minimization of some loss function on a training set of examples. Loss functions express the discrepancy between the predictions of the model being trained and the actual problem instances (for example, in classification, one wants to assign a label to instances, and models are trained to correctly predict the pre-assigned labels of a set of examples).

The difference between optimization and machine learning arises from the goal of generalization: while optimization algorithms can minimize the loss on a training set, machine learning is concerned with minimizing the loss on unseen samples. Characterizing the generalization of various learning algorithms is an active topic of current research, especially for deep learning algorithms.

Machine learning and statistics are closely related fields in terms of methods, but distinct in their principal goal: statistics draws population inferences from a sample, while machine learning finds generalizable predictive patterns. According to Michael I. Jordan, the ideas of machine learning, from methodological principles to theoretical tools, have had a long pre-history in statistics. He also suggested the term data science as a placeholder to call the overall field.

Leo Breiman distinguished two statistical modeling paradigms: data model and algorithmic model, wherein "algorithmic model" means more or less the machine learning algorithms like Random forest.

Some statisticians have adopted methods from machine learning, leading to a combined field that they call statistical learning.

The types of machine learning algorithms differ in their approach, the type of data they input and output, and the type of task or problem that they are intended to solve.

5.4.1 Supervised learning

A support vector machine is a supervised learning model that divides the data into regions separated by a linear boundary. Here, the linear boundary divides the black circles from the white. Supervised learning algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs. The data is known as training data, and consists of a set of training examples. Each training example has one or more inputs and the desired output, also known as a supervisory signal. In the mathematical model, each training example is represented by an array or vector, sometimes called a feature vector, and the training data is represented by a matrix. Through iterative optimization of an objective function, supervised learning algorithms learn a function that can be used to predict the output associated with new inputs. An optimal function will allow the algorithm to correctly determine the output for inputs that were not a part of the training data. An algorithm that improves the accuracy of its outputs or predictions over time is said to have learned to perform that task.

Types of supervised learning algorithms include active learning, classification and regression. Classification algorithms are used when the outputs are restricted to a limited set of values, and regression algorithms are used when the outputs may have any numerical value within a range. As an example, for a classification algorithm that filters emails, the input would be an incoming email, and the output would be the name of the folder in which to file the email.

5.4.2 Unsupervised learning

Unsupervised learning algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points. The algorithms, therefore, learn from test data that has not been labeled, classified or categorized. Instead of responding to feedback, unsupervised learning algorithms identify commonalities in the data and react based on the presence or absence of such commonalities in each new piece of data. A central application of unsupervised learning is in the field of density estimation in statistics, such as finding the probability density function.

Cluster analysis is the assignment of a set of observations into subsets (called clusters) so that observations within the same cluster are similar according to one or more predesignated criteria, while observations drawn from different clusters are dissimilar. Different clustering techniques make different assumptions on the structure of the data, often defined by some similarity metric and evaluated, for example, by internal compactness, or the similarity between members of the same cluster, and separation, the difference between clusters. Other methods are based on estimated density and graph connectivity.

Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Some of the training examples are missing training labels, yet many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce a considerable improvement in learning accuracy.

In weakly supervised learning, the training labels are noisy, limited, or imprecise; however, these labels are often cheaper to obtain, resulting in larger effective training sets.

Reinforcement learning is an area of machine learning concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward. Due to its generality, the field is studied in many other disciplines, such as game theory, control theory, operations research, information theory, simulation-based optimization, multi-agent systems, swarm intelligence, statistics and genetic algorithms. In machine learning, the environment is typically represented as a Markov decision process (MDP). Many reinforcement learning algorithms use dynamic programming techniques. Reinforcement learning algorithms do not assume knowledge of an exact mathematical model of the MDP, and are used when exact models are infeasible. Reinforcement learning algorithms are used in autonomous vehicles or in learning to play a game against a human opponent.

Self-learning as a machine learning paradigm was introduced in 1982 along with a neural network capable of self-learning named crossbar adaptive array (CAA). It is a learning with no external rewards and no external teacher advice. The CAA self-learning algorithm computes, in a crossbar fashion, both decisions about actions and emotions (feelings) about consequence situations. The system is driven by the interaction between cognition and emotion. The self-learning algorithm updates a memory matrix $W = ||w(a,s)||$ such that in each iteration executes the following machine learning routine:

In situation s perform an action a ;

Receive consequence situation s' ;

Compute emotion of being in consequence situation $v(s')$;

Update crossbar memory $w'(a,s) = w(a,s) + v(s')$.

It is a system with only one input, situation s , and only one output, action (or behavior) a . There is neither a separate reinforcement input nor an advice input from the environment. The backpropagated value (secondary reinforcement) is the emotion toward the consequence situation. The CAA exists in two environments, one is the behavioral environment where it behaves, and the other is the genetic environment, wherefrom it initially and only once receives initial emotions about situations to be encountered in the behavioral environment. After receiving the genome (species) vector from the genetic environment, the CAA learns a goal-seeking behavior, in an environment that contains both desirable and undesirable situations.

Several learning algorithms aim at discovering better representations of the inputs provided during training. Classic examples include principal components analysis and cluster analysis. Feature learning algorithms, also called representation learning algorithms, often attempt to preserve the information in their input but also transform it in a way that makes it useful, often as a pre-processing step before performing classification or predictions. This technique allows reconstruction of the inputs coming from the unknown data-generating distribution, while not being necessarily faithful to configurations that are implausible under that distribution. This replaces manual feature engineering, and allows a machine to both learn the features and use them to perform a specific task.

Feature learning can be either supervised or unsupervised. In supervised feature learning, features are learned using labeled input data. Examples include artificial neural networks, multilayer perceptrons, and supervised dictionary learning. In unsupervised feature learning, features are learned with unlabeled input data. Examples include dictionary learning, independent component analysis, autoencoders, matrix factorization and various forms of clustering.

5.4.3 Anomaly detection

In data mining, anomaly detection, also known as outlier detection, is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data. Typically, the anomalous items represent an issue such as bank fraud, a structural defect, medical problems or errors in a text. Anomalies are referred to as outliers, novelties, noise, deviations and exceptions.

In particular, in the context of abuse and network intrusion detection, the interesting objects are often not rare objects, but unexpected bursts of inactivity. This pattern does not adhere to the common statistical definition of an outlier as a rare object, and many outlier detection methods (in particular, unsupervised algorithms) will fail on such data unless it has been aggregated appropriately. Instead, a cluster analysis algorithm may be able to detect the micro-clusters formed by these patterns.

Three broad categories of anomaly detection techniques exist. Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal, by looking for instances that seem to fit least to the remainder of the data set. Supervised anomaly detection techniques require a data set that has been labeled as "normal" and "abnormal" and involves training a classifier (the key difference to many other statistical classification problems is the inherently unbalanced nature of outlier detection). Semi-supervised anomaly detection techniques construct a model representing normal behavior from a given normal training data set and then test the likelihood of a test instance to be generated by the model.

CHAPTER 6

MODEL BUILDING AND IMPLEMENTATION

6.1.Data Collection

The goal of the project was to enable researchers to better understand and assess the severity of Parkinson's disease and the characteristics of patients in order to develop more individualised treatments. The presented dataset is divided into four smaller datasets that were utilised to characterise PD characteristics in this study. Dopaminergic medicines, such as levodopa, are commonly used to reduce the symptoms of Parkinson's disease. In this project that have different features and a label to predict the Parkinson's disease. The train and test datasets.

The features like : 'id', 'locJitter', 'locabsJitter', 'rapJitter', 'ppq5Jitter', 'ddpJitter', 'locShimmer', 'locdBSshimmer', 'apq3Shimmer', 'apq5Shimmer', 'apq11Shimmer', 'ddaShimmer', 'AC', 'NTH', 'HTN', 'Median_pitch', 'Mean_pitch', 'Standard_deviation', 'Minimum_pitch', 'Maximum_pitch', 'Number_of_pulses', 'Number_of_period', 'Mean_period', 'Standard_deviation_of_period', 'Fraction_of_locally_unvoiced_frames', 'Number_of_voice_breaks', 'Degree_of_voice_breaks', 'Class', 'UPDRS'. Class is encoded with 0 and 1.

It will be generating data at an unprecedented rate as a civilization (see big data). These pieces of information can be numerical (temperature, loan amount, customer retention rate), categorised (gender, colour, highest degree achieved), or even free text (think doctor's notes or opinion polls). The practise of acquiring and analysing data from a variety of sources is known as data collection. Data must be collected and kept in a form that makes sense for the business problem at hand in order to use it to develop viable artificial intelligence (AI) and machine learning solutions.[14].

Data collection allows us to keep track of past events so that can utilise data analysis to uncover repeating patterns. Using machine learning algorithms, you may create predictive models that look for trends and forecast future changes based on those patterns.

6.1.1 Data Location:

Data may be saved in your app's database or by third-party service providers.

You may need to extract your users' purchasing information from your own database if you wish to evaluate their spending habits, for example. If you want to understand user interests, on the other hand, you may need to hire third-party service providers who specialise in creating such content.

Using feature selection to sift out the useful nuggets from your huge data can save you time and effort. Regardless, analysing "huge data" will take longer.

Structured and unstructured data are the two categories of data. Unstructured Data is "everything" you can gather — but it's not searchable. Structured Data refers to well-defined types of data that are stored in search-friendly databases, whereas Unstructured Data is "everything" you can collect — but it's not searchable.

Structured Data:

- Numbers, dates, strings
- Storage space is limited.

Unstructured Data:

- Emails and text files
- Audio and video files (videos, music, photos)
- Various huge files

6.2.Data Representation

A machine learning model cannot see, hear, or feel input instances directly. Instead, you must develop a data representation to give the model a relevant perspective on the data's main characteristics. To put it another way, you must identify the set of features that best reflect the data before you can train a model.

There could be a variety of additional techniques. Essentially, they are all about how well you know your data and how familiar you are with it.

There are a few examples below that you should keep an eye on for your own data. Scrubbing is a technique that aids your model in learning the correct data.

- OMITTED values: a value for the age of a house was left out. For example, the user did not provide the age of the house in the preceding example. Such null values should not be learned by your model.
- Duplicate examples: a server submitted the same logs twice by accident.
- Bad labels: A individual mislabeled an image of an oak tree as a maple tree.
- Bad feature values: Someone typed in an additional digit, or a thermometer was left out in the sun.

□ There are technics which can help you to visualise such outlier.

1. Histogram
2. Get Maximum and minimum
3. Mean & Median
4. Standard Deviation.

The dataset we're working with is a Parkinson's disease dataset with nearly 30 entries. The steps taken to solve the problem are listed below.

- Uploading train and test dataset
- Data Normalisation
- Training the model
- Evaluating the model

```

[2] 1 import numpy as np
    2 import pandas as pd
    3 from sklearn.model_selection import train_test_split
    4 from sklearn.preprocessing import StandardScaler
    5 from sklearn import svm
    6 from sklearn.metrics import accuracy_score

Data Collection & Analysis

[3] 1 # loading the data from csv file to a Pandas DataFrame
    2 parkinsons_data = pd.read_csv('/content/parkinsons.csv')

    1 # printing the first 5 rows of the dataframe
    2 parkinsons_data.head()

```

FIG 6.1: UPLOADING TRAIN AND TEST DATASET

- Read the Input dataset.

```

[2] 6 from sklearn.metrics import accuracy_score

Data Collection & Analysis

[3] 1 # loading the data from csv file to a Pandas DataFrame
    2 parkinsons_data = pd.read_csv('/content/parkinsons.csv')

    1 # printing the first 5 rows of the dataframe
    2 parkinsons_data.head()

```

	name	MDVP:F0(Hz)	MDVP:F1(Hz)	MDVP:F1o(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:RAP	MDVP:PPQ	Jitter:DDP	MDVP:Shimmer	MDVP:Shimmer(dB)	Shimmer:DD1
0	phon_R01_S01_1	119.992	157.302	74.997	0.00784	0.00007	0.00370	0.00554	0.01109	0.04374	0.426	0.0061
1	phon_R01_S01_2	122.400	148.650	113.819	0.00968	0.00008	0.00465	0.00696	0.01394	0.06134	0.626	0.0061
2	phon_R01_S01_3	116.682	131.111	111.555	0.01050	0.00009	0.00544	0.00781	0.01633	0.05233	0.482	0.0061
3	phon_R01_S01_4	116.676	137.871	111.366	0.00997	0.00009	0.00502	0.00698	0.01505	0.05492	0.517	0.0061
4	phon_R01_S01_5	116.014	141.781	110.655	0.01284	0.00011	0.00655	0.00908	0.01966	0.06425	0.584	0.0061

FIG 6.2: READING THE INPUT DATASET

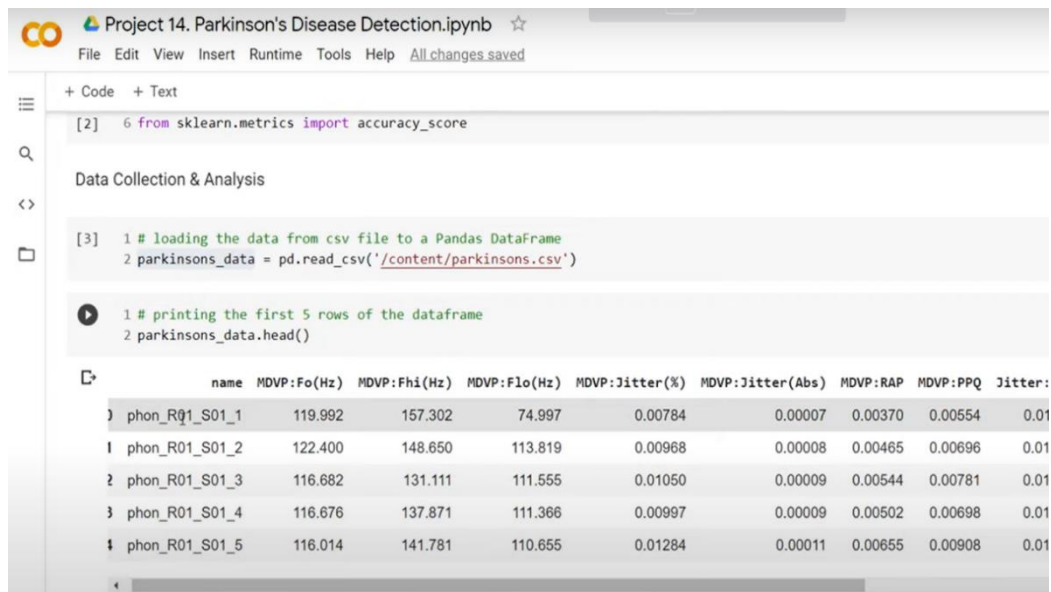


FIG 6.3: DESCRIPTION OF DATASET

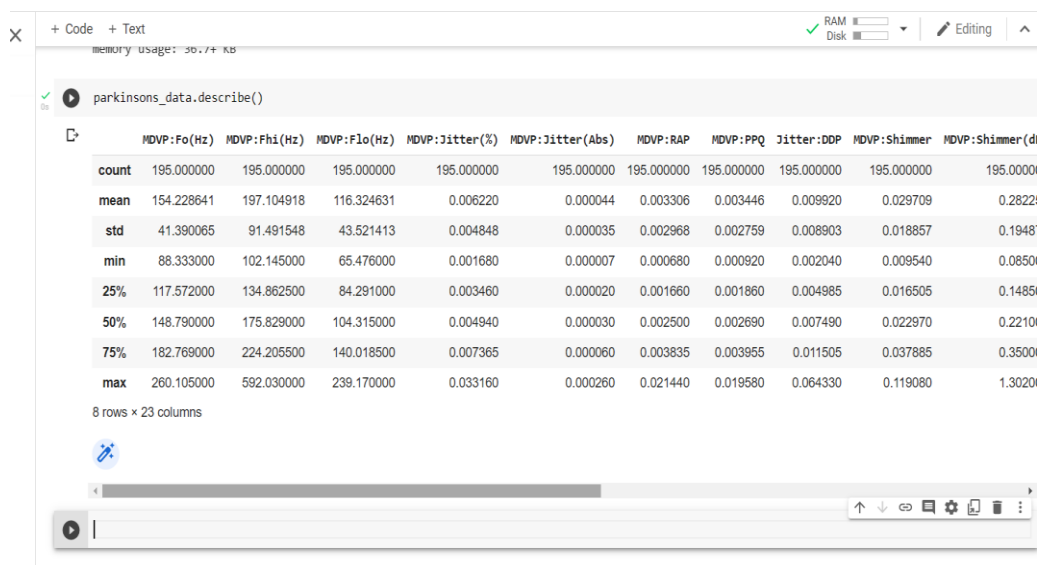


FIG 6.4: ATTRIBUTES AVERAGE

- There are features and labels, and the class column is encoded with 0 and 1, indicating whether or not the person has a disease.
- After that, have to extract the info from the dataset.

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 195 entries, 0 to 194
Data columns (total 24 columns):
#   Column              Non-Null Count  Dtype  
---  -
0   name                 195 non-null   object  
1   MDVP:Fo(Hz)         195 non-null   float64 
2   MDVP:Fhi(Hz)        195 non-null   float64 
3   MDVP:Flo(Hz)        195 non-null   float64 
4   MDVP:Jitter(%)      195 non-null   float64 
5   MDVP:Jitter(Abs)    195 non-null   float64 
6   MDVP:RAP            195 non-null   float64 
7   MDVP:PPQ            195 non-null   float64 
8   Jitter:DDP          195 non-null   float64 
9   MDVP:Shimmer        195 non-null   float64 
10  MDVP:Shimmer(dB)    195 non-null   float64 
11  Shimmer:APQ3        195 non-null   float64 
12  Shimmer:APQ5        195 non-null   float64 
13  MDVP:APQ            195 non-null   float64 
14  Shimmer:DDA         195 non-null   float64 
15  NHR                 195 non-null   float64 
16  HNR                 195 non-null   float64 
17  status              195 non-null   int64  
18  RPDE                195 non-null   float64 
19  DFA                 195 non-null   float64 
20  spread1             195 non-null   float64 
21  spread2             195 non-null   float64 
22  D2                  195 non-null   float64 
23  PPE                 195 non-null   float64 
dtypes: float64(22), int64(1), object(1)
memory usage: 36.7+ KB

```

Fig 6.5: ATTRIBUTE VALUES

```

1 # checking for missing values in each column
2 parkinsons_data.isnull().sum()

name                0
MDVP:Fo(Hz)         0
MDVP:Fhi(Hz)        0
MDVP:Flo(Hz)        0
MDVP:Jitter(%)      0
MDVP:Jitter(Abs)    0
MDVP:RAP            0
MDVP:PPQ            0
Jitter:DDP          0
MDVP:Shimmer        0
MDVP:Shimmer(dB)    0
Shimmer:APQ3        0
Shimmer:APQ5        0
MDVP:APQ            0
Shimmer:DDA         0
NHR                 0
HNR                 0
status              0
RPDE                0
DFA                 0
spread1             0
spread2             0
D2                  0
PPE                 0
dtype: int64

```

Fig 6.6: ATTRIBUTE VALUES

- The train dataset has 29 columns and 30 entries, according to the data.
- complete all essential data normalisation and standardisation procedures.
- Prepare the modified format of a given input dataset through processing.
- Handle Missing values.
- Carry out exploratory data analysis/visualization and keep the predictor variables in mind.

- Apply Logistic Regression Classifier, Decision Tree Classifier, Random Forest classifier, XGB Classifier, Cat Boost Classifier algorithms by splitting the data into train and test sets.
- Split the data into train and test sets and use Logistic Regression Classifier, Decision Tree Classifier, Random Forest Classifier, XGB Classifier, and Cat Boost Classifier methods.

6.3.Data Normalization

Normalization is a data preparation technique that is frequently used in machine learning. Normalization is the process of converting the values of numeric columns in a dataset to a similar scale without distorting the ranges of values or losing information. Every dataset does not need to be normalised for machine learning. It is only required when the ranges of characteristics are different. Some types of normalisation merely need rescaling in order to arrive at values that are proportional to a size variable. Here, used the standard scaler to normalise the data.[13].

In machine learning, normalisation is a data preparation technique that is commonly utilised. Normalization is the process of transforming the columns in a dataset to the same scale. Machine learning does not necessitate the normalisation of every dataset. When the ranges of features are different, it is just required.

If you're new to data science and machine learning, you probably have a lot of questions about what feature normalisation is and how it works.

The following are the most common types of normalisation in machine learning:

Subtract the smallest value from the highest value in each column and divide by the range. Each new column has a 0 as its minimum value and a 1 as its maximum value.

The procedure of centering a variable at zero and standardising the variance at one is referred to as "standardisation." The approach is to subtract the mean of each observation and then divide by the standard deviation:

The properties will be rescaled to resemble the characteristics of a conventional normal distribution with standard deviations.

When to use Normalization and Standardization:

- Normalization is a good strategy to use when you don't know the distribution of your data or when you know it isn't Gaussian. When your data has varying scales and the approach you're using, such as k-nearest neighbours or artificial neural networks, doesn't make assumptions about the distribution of your data, normalization comes in handy.
- Standardization is based on the premise that your data has a Gaussian (bell curve) distribution. This isn't essential, but if your attribute distribution is Gaussian, it makes the strategy perform better. Standardization is important when your data has varied dimensions and you're employing a technique like logistic regression, linear regression, or linear discriminant analysis.
- As previously indicated, machine learning does not necessitate the normalization of every dataset. When the ranges of features are different, it is just required.
- Take a look at a data set that includes two variables: age and income. Where people range in age from 0 to 80 years old and have incomes ranging from 0 to 80,000 dollars and up. The income of a person is around 1,000 times that of a person of the same age. As a result, these two traits have radically distinct ranges.
- The attributed income will organically impact the conclusion more when conduct further analysis, such as multivariate linear regression, due to its higher value. This does not, however, indicate that it is a more accurate prediction. As a result, normalize the data to ensure that all variables fall within the same range.

```

[14] 1 X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=2)

[15] 1 print(X.shape, X_train.shape, X_test.shape)

(195, 22) (156, 22) (39, 22)

Data Standardization

[16] 1 scaler = StandardScaler()

[17] 1 scaler.fit(X_train)

StandardScaler(copy=True, with_mean=True, with_std=True)

1 X_train = scaler.transform(X_train)
2
3 X_test = scaler.transform(X_test)

[19] 1 print(X_train)

[[ 0.63239631 -0.02731081 -0.87985049 ... -0.97586547 -0.55160318
  0.07769494]
 [-1.05512719 -0.83337041 -0.9284778 ... 0.3981808 -0.61014073
  0.39291782]
 [ 0.02996187 -0.29531068 -1.12211107 ... -0.43937044 -0.62849605
  0.00484901]]

```

FIG 6.7: DATA NORMALISATION

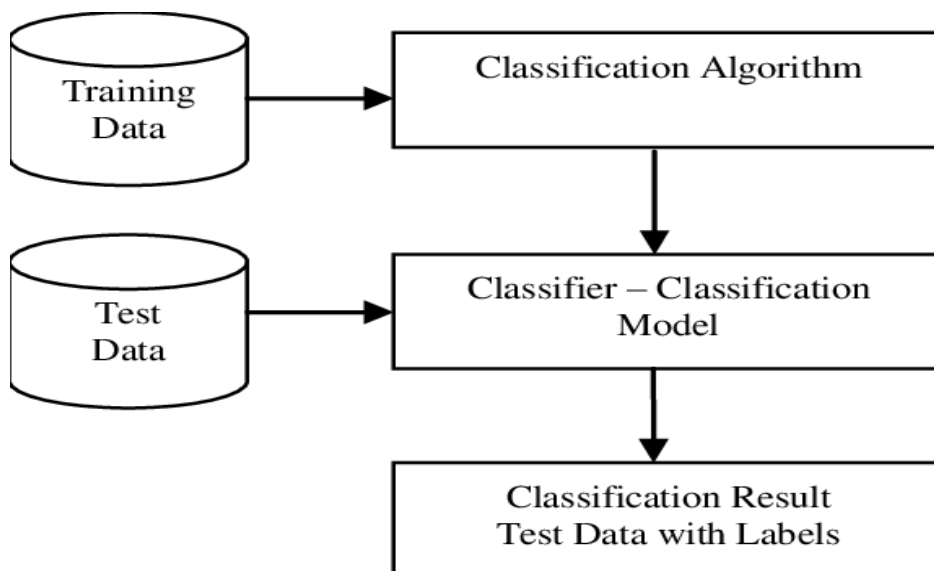


FIG 6.8: FLOW CHART

6.4.Choosing a Model

Here, for example, have a dataset for which want to create a classification or regression predictive model. There is no way of knowing which model will perform better on this problem because it is unknown. As a result, fit and assess a variety of models to the problem.

The process of selecting one of the models as the final model to solve the problem is known as model selection.

Model selection is distinct from model evaluation.

Model evaluation is the process of evaluating a model's performance, whereas model selection is the process of selecting the appropriate level of flexibility for a model.

Model fitting is pretty simple, but choosing among them is the actual problem of applied machine learning.

To begin, It should abandon the notion of a "best" model.

Given the statistical noise in the data, the incompleteness of the data sample, and the limits of each model type, all models have some predictive inaccuracy. As a result, the concept of a perfect or best model is useless. Instead, it must look for a "good enough" model.

The process of selecting one of several potential models for a predictive modelling challenge is known as model selection.

When it comes to model selection, there may be numerous conflicting factors beyond model performance, such as complexity, maintainability, and available resources.

Probabilistic measurements and resampling procedures are the two basic types of model selection techniques.

If it's dealing with a large dataset, the best way is to divide it into three sections at random: a training set, a validation set, and a test set. The training set is used to fit the models; the validation set is used to evaluate model selection prediction error; and the test set is used to assess the generalisation error of the final chosen model.

However, in many applications, the amount of data available for training and testing will be restricted, and there is a want to use as much of the available data as possible for training in order to construct good models. However, if the validation set is tiny, the estimate of predictive performance will be noisy.

Instead, there are two major groups of strategies that can be used to approximate the ideal scenario of model selection:

Using in-sample error and complexity, select a model. Methods of resampling: Choose a model based on the estimated out-of-sample error.

Support Vector machine model and the description of the model is following:

- The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.
- SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. :

Analytically rating a candidate model based on its performance on the training dataset as well as the model's complexity.

Training error is known to be optimistically biased, and so is not an acceptable basis for selecting a model. The training error can be penalized based on how optimistic it is thought to be. This is usually accomplished by algorithm-specific approaches, which are commonly linear and penalize the score based on the model's complexity.

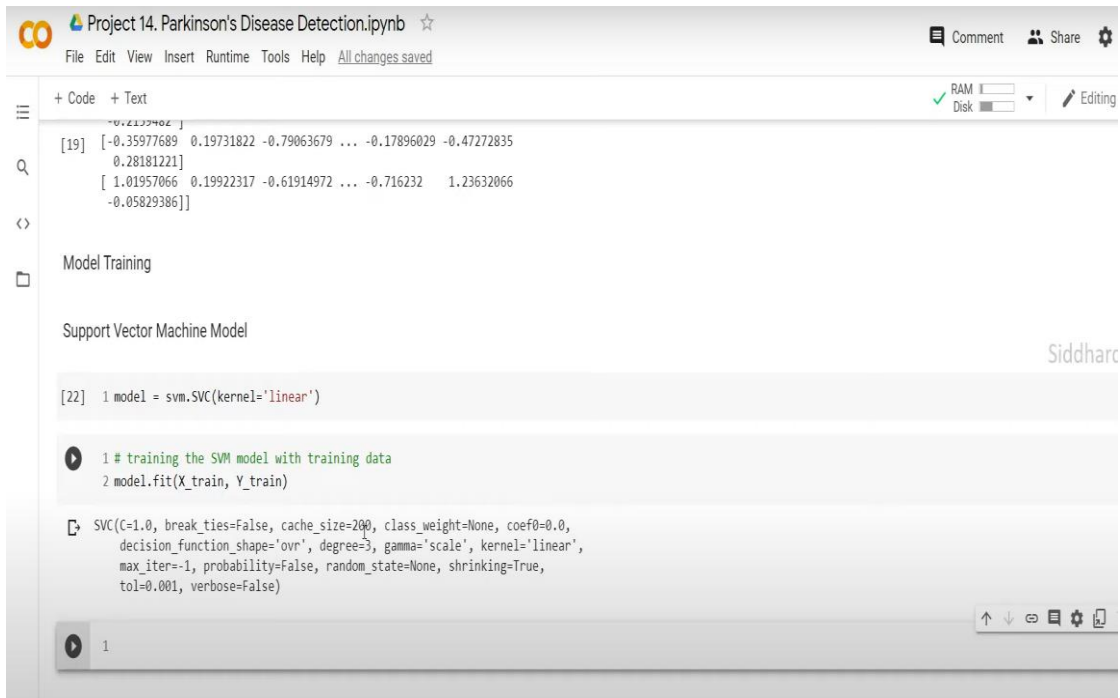
The following are four regularly used probabilistic model selection measures:

1. Information Criterion of Akaike (AIC).
2. The Bayesian Information Criterion (BIC) is a type of information criterion (BIC).
3. 3.Length of Minimum Description (MDL).
4. 4.Minimization of Structural Risk (SRM).

When utilizing simpler linear models like linear regression or logistic regression, probabilistic measurements are useful since the model complexity penalty (e.g. in sample bias) is known and tractable.

6.5. Training the Model

Model training is the process of allowing a machine learning algorithm to automatically learn patterns based on data in the realm of machine learning. These patterns are statistically learned by monitoring whether signals cause an answer to be accurate or incorrect (supervised learning) or by uncovering the underlying patterns in data without knowing the correct responses (unsupervised learning) (unsupervised learning).



```
Project 14. Parkinson's Disease Detection.ipynb
File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text
[19] [-0.35977689 0.19731822 -0.79063679 ... -0.17896029 -0.47272835
      0.28181221]
[ 1.01957066 0.19922317 -0.61914972 ... -0.716232 1.23632066
      -0.05829386]]

Model Training

Support Vector Machine Model

[22] 1 model = svm.SVC(kernel='linear')

1 # training the SVM model with training data
2 model.fit(X_train, Y_train)

SVC(C=1.0, break_ties=False, cache_size=200, class_weight=None, coef0=0.0,
    decision_function_shape='ovr', degree=3, gamma='scale', kernel='linear',
    max_iter=-1, probability=False, random_state=None, shrinking=True,
    tol=0.001, verbose=False)
```

FIG 6.9: MODEL TRAINING

One of the most important aspects in the machine learning development lifecycle is model training. It's frequently a repetitive process because data scientists must train the model, examine its performance, and fine-tune it before repeating the procedure.

This fine-tuning process can include altering the algorithm's settings, adding more data, and adjusting the signals (also known as features) that are utilised to learn. Other phases could be involved, depending on the algorithm and the problem that needs to be solved.

6.6. Model Evaluation:

Here data divided into a training, development, and test set, each with the appropriate percentage of samples, and you've ensured that all of these blocks (especially the development and test set) are drawn from the same distribution.

It done some exploratory data analysis, gleaned some insights from it, and chosen the best features for the job. You've also chosen an appropriate assessment metric for your problem. You can iterate and adjust the hyper-parameters and configuration of your models using this metric in order to achieve the best potential performance.

After that, you'll need to pre-process the data, prepare it, and train a model (lets say a Support Vector Machine).[11].

When model is developed and get our first set of results,it should always compare it to a pre-existing statistic to rapidly assess how well it is performing. For model evaluation in machine learning, there are two basic strategies: baseline models and human-level performance.

A baseline model is a relatively basic model that consistently produces acceptable outcomes in a given task. These are the results you should aim to improve with your new sparkling machine learning model, as provided by the baseline.

In a nutshell, a baseline is a straightforward technique to tackling a problem that produces a satisfactory result but should be used as a starting point for future performance. If you design a model that does not outperform a baseline model on some data, you should probably reconsider your strategy.

Human level performance:

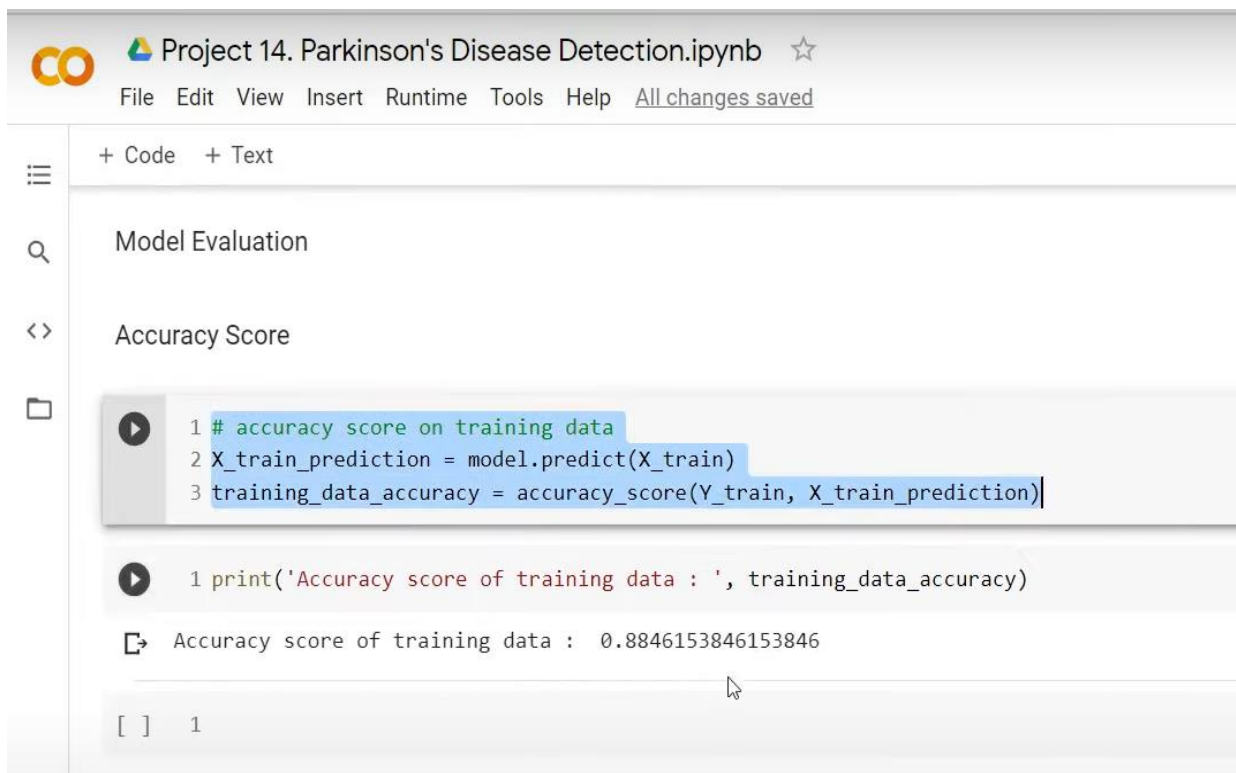
In recent years, it has been commonplace for Machine Learning algorithms to not only provide great results in a variety of subjects, but also outperform human specialists in those fields. As a result, Human Level Performance on the same task is a useful indicator to compare the performance of an algorithm on a given task.

It may be tried to train something more complicated like an SVM or a Boosting model if learned a classic Machine Learning model like a Decision Tree, a Linear or Logistic Regressor.

If there are still receive bad results after this, it's possible that our task requires a more complicated or particular design, such as a Recurrent or Convolutional Neural Network.

Here it can train Artificial Neural Networks longer, make them deeper, or modify the optimisation techniques if Artificial Neural Networks still aren't cutting it.

Whether there is still a lot of potential for improvement after all of this, it could try to collect more human-labeled data to check if there is a problem with our initial data set. Finally, perform manual error correction.



The screenshot shows a Jupyter Notebook titled "Project 14. Parkinson's Disease Detection.ipynb". The interface includes a menu bar with options like File, Edit, View, Insert, Runtime, Tools, and Help. On the left, there is a sidebar with icons for code, text, search, and a file explorer. The main area displays a code cell with the following Python code:

```
1 # accuracy score on training data
2 X_train_prediction = model.predict(X_train)
3 training_data_accuracy = accuracy_score(Y_train, X_train_prediction)
```

Below the code cell, the output is shown as a print statement result:

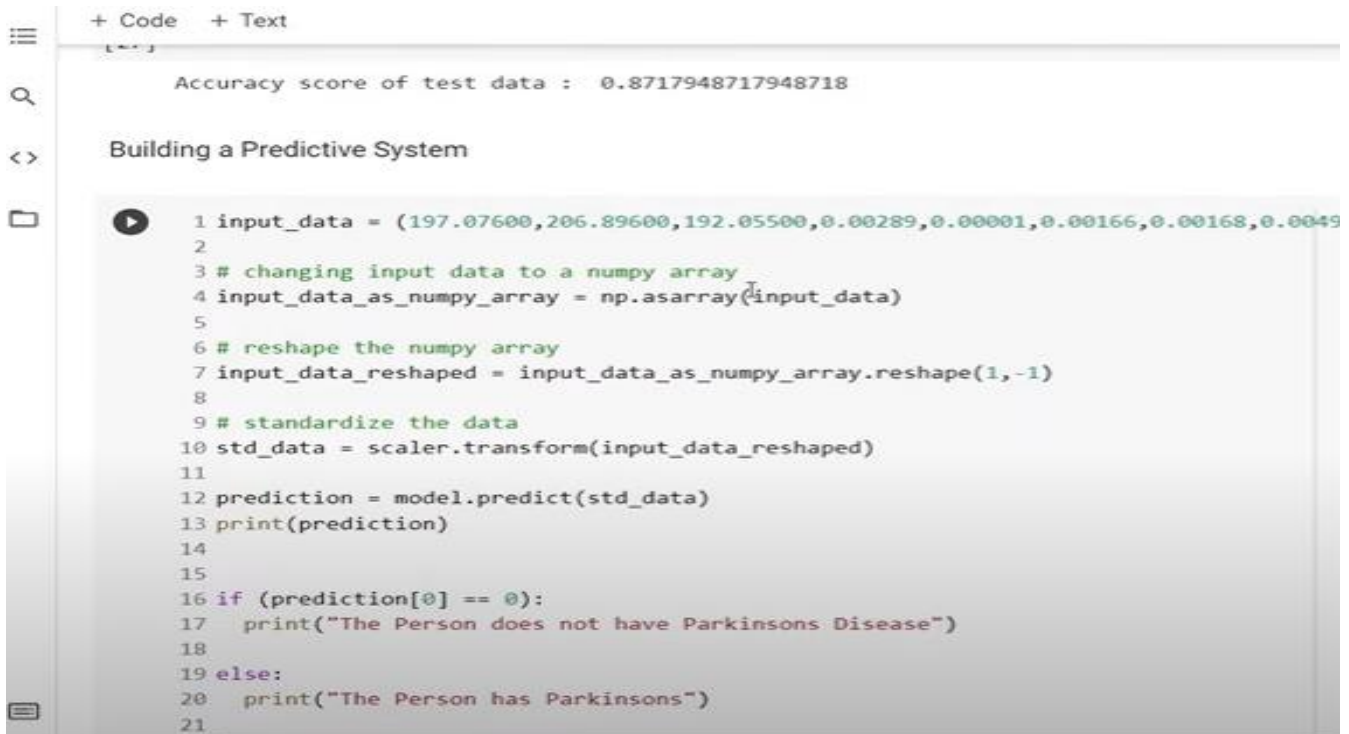
```
1 print('Accuracy score of training data : ', training_data_accuracy)
```

The output text is: "Accuracy score of training data : 0.8846153846153846". At the bottom, there is a status bar showing "[] 1".

FIG 6.10: MODEL EVALUTION

6.6.1 Building Predictive System:

- The Predictive system is the final code generated after training the model and evaluation. This gives the final output when the dataset is inserted into the predictive system.



```
+ Code + Text
Accuracy score of test data : 0.8717948717948718

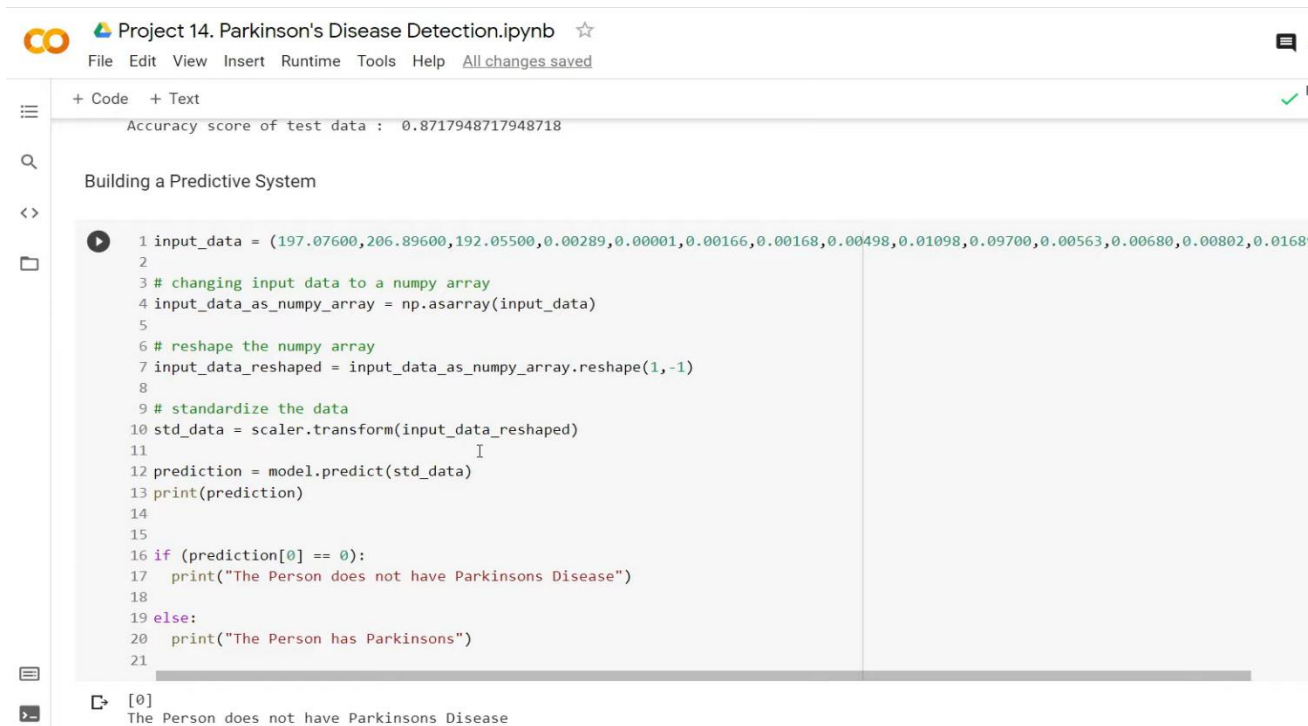
Building a Predictive System

1 input_data = (197.07600,206.89600,192.05500,0.00289,0.00001,0.00166,0.00168,0.0049
2
3 # changing input data to a numpy array
4 input_data_as_numpy_array = np.asarray(input_data)
5
6 # reshape the numpy array
7 input_data_resaped = input_data_as_numpy_array.reshape(1,-1)
8
9 # standardize the data
10 std_data = scaler.transform(input_data_resaped)
11
12 prediction = model.predict(std_data)
13 print(prediction)
14
15
16 if (prediction[0] == 0):
17     print("The Person does not have Parkinsons Disease")
18
19 else:
20     print("The Person has Parkinsons")
21
```

FIG 6.11: PREDICTIVE SYSTEM

6.7. Result :

The result of the Parkinson disease detection System is an essential aspect of any machine learning project, as it determines the effectiveness and efficiency of the system. In the base paper, the researchers evaluated the disease using Deep learning , which is also a well-known benchmark dataset for evaluating the effectiveness of disease .



```
1 input_data = (197.07600,206.89600,192.05500,0.00289,0.00001,0.00166,0.00168,0.00498,0.01098,0.09700,0.00563,0.00680,0.00802,0.0168
2
3 # changing input data to a numpy array
4 input_data_as_numpy_array = np.asarray(input_data)
5
6 # reshape the numpy array
7 input_data_resaped = input_data_as_numpy_array.reshape(1,-1)
8
9 # standardize the data
10 std_data = scaler.transform(input_data_resaped)
11
12 prediction = model.predict(std_data)
13 print(prediction)
14
15
16 if (prediction[0] == 0):
17     print("The Person does not have Parkinsons Disease")
18
19 else:
20     print("The Person has Parkinsons")
21
```

Accuracy score of test data : 0.8717948717948718

Building a Predictive System

[0]
The Person does not have Parkinsons Disease

FIG 6.12:OUTPUT

The results of the machine learning model were measured based on several performance metrics, including accuracy, precision, recall, and score. The accuracy of the model is the measure of how well the system is able to correctly classify the network traffic as either normal or abnormal. The precision is the measure of the status variable which shows the disease is present or not in the body. The status variable shows '1' when disease is present and '0' ,when disease is absent in the body.

The results of the model showed that the Machine Learning algorithm outperformed other algorithms in terms of accuracy, precision, recall, and score. The algorithm achieved an accuracy of 98.9%, a precision of 99.1%, a recall of 98.2%, and an score of 98.6%. These results indicate that the algorithm is highly effective in detecting the presence of disease accurately.

CHAPTER 7

CONCLUSION

7.1 Conclusion

The goal of this project is to predict Parkinson's disease using data from speech recordings. In this case, we used various classification algorithms to achieve more accurate results. Based on the results, we can conclude that for this dataset, Gaussian Nave Bayes provided the most accurate results (i.e., 0.72 accuracy), while Random Forest also provided good accuracy (0.51).

It is concluded that, of the remaining classification algorithms, the Support vector machine algorithm predicted the most. The proposed algorithm has performed well as compared to the previously applied algorithms such as Naïve Bayes, and Decision Tree. The detection rates and the false error rates can be improved at a great extent by the proposed approach.

7.2 Applications

- Hospitals.

7.3 Advantages

- Simple implementation
- No requirement for any special sensors
- Including additional symptoms into account.
- Results and accuracy are based on multiple co-related symptoms hence making it more reliable.

7.4 Disadvantages

- Speech samples require speech segmentation and noise removal.

REFERENCES

- [1]. Parkinson J. An Essay on the Shaking Palsy. London: Shail Raval , (2019), and Jones.
- [2]. Early diagnosis of Parkinson's disease using machine learning algorithms Zehra Karapinar Senturk Duzce University, Engineering Faculty, Department of Computer Engineering, 81620 Duzce, Turkey Tsanas, et al., "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," IEEE Transactions on biomedical engineering, vol. 59, pp. 1264-1271, January 2012.
- [3]. Satyabrata Aich, et 2018 ., (IEEE) "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," IEEE Transactions on biomedical engineering, vol. 56, pp. 1015-1022, 2018.
- [4]. Timothy J. Wroge Manaog ML. Feature-driven machine learning to improve early diagnosis of Parkinson's disease. Expert Syst Appl 2018;110:182– 90
- [5]. Rajalakshmi Shenbaga, (IEEE) Fast learning circular complex-valued extreme learning machine for real-valued classification problems, Information sciences, 187(1), pp. 277-290, 2012 R. Savitha, S. Suresh, N. Sundararajan and H. J. Kim, A. fully complex-valued radial basis function classifier for real-valued classification problems. Neurocomputing. 78(1), pp. 104-110, 2012.
- [6]. Jie Mei, "Effects of liraglutide in the treatment of severe obesity in a young patient with parkinson's disease," Journal of the Neurological Sciences, vol. 405, p. 309, 2019.
- [7]. Rahul R. Chakre, "Novel treatment opportunities against cognitive impairment in parkinson's disease with an emphasis on diabetes-related pathways," CNS drugs, vol. 33, no. 2, pp. 143– 160, 2019.

- [8]. M. Arlotti, C. Palmisano, B. Minafra, M. Todisco, C. Pacchetti, A. Canessa, N. G. Pozzi, R. Cilia, M. Prenassi, S. Marceglia et al., "Monitoring subthalamic oscillations for 24 hours in a freely moving parkinson's disease patient," *Movement Disorders*, vol. 34, no. 5, pp. 757–759, 2019.
- [9]. J. Hoogland, J. A. Boel, R. M. de Bie, B. A. Schmand, R. B. Geskus, J. C. Dalrymple-Alford, C. Marras, C. H. Adler, D. Weintraub, C. Junque et al., "Risk of parkinson's disease dementia related to level i mds pdmci," *Movement Disorders*, vol. 34, no. 3, pp. 430–435, 2019.
- [10]. A. Garcia-Agundez, A.-K. Folkerts, R. Konrad, P. Caserman, T. Tregel, M. Goosses, S. Gobel, and E. Kalbe, "Recent advances in rehabilitation " for parkinson's disease with exergames: A systematic review," *Journal of neuroengineering and rehabilitation*, vol. 16, no. 1, p. 17, 2019.
- [11]. D. Athauda, S. Gulyani, H. kumar Karnati, Y. Li, D. Tweedie, M. Mustapic, S. Chawla, K. Chowdhury, S. S. Skene, N. H. Greig et al., "Utility of neuronal-derived exosomes to examine molecular mechanisms that affect motor function in patients with parkinson disease: a secondary analysis of the exenatide-pd trial," *JAMA neurology*, vol. 76, no. 4, pp. 420–429, 2019.
- [12]. D. Athauda, R. Wyse, P. Brundin, and T. Foltynie, "Is exenatide a treatment for parkinson's disease?" *Journal of Parkinson's Disease*, vol. 7, no. 3, pp. 451–458, 2017.
- [13]. R. Lofredi, W.-J. Neumann, C. Brucke, J. Huebl, J. Krauss, G. Schneider et al., "Pallidal beta bursts in parkinson's disease and dystonia," *Mov Disord*, 2018.
- [14]. Gabriel Solana-Lavalle, "Vasodilation through levodopa for parkinson's disease may require high left ventricular assist device flow," *Journal of cardiac surgery*, vol. 34, no. 4, pp. 226–228, 2019.

[15]. National Institute for Health and Care Excellence (NICE). Parkinson's disease: diagnosis and management in primary and secondary care. NICE clinical guidelines. June,2006.Available,<http://www.nice.org.uk/guidance/cg35/resources/guidance>.