COMPARATIVE ANALYSIS OF GUI BASED PREDICTION OF PARKINSON DISEASE USING MACHINE LEARNING APPROACH

Mr. S.RAMAKRISHNAN [1], R.SURYA[2], R.VISHAL[3] , M.SAMSON[4], R.SHARATH KUMAR [5]

[1]Head Of The Department, Dept. of IT, Jeppiaar SRR Engineering College, Chennai, Tamil Nadu
[2,3,4,5]B.TECH., Dept. of IT, Jeppiaar SRR Engineering College, Chennai, Tamil Nadu

---

Abstract - *Parkinson's disease is the most prevalent neurodegenerative disorder affecting more than 10 million people worldwide. There is no single test which can be administered for diagnosing Parkinson's disease. Because of these difficulties, to investigate a machine learning approach to accurately diagnose Parkinson's, using a given dataset. To prevent this problem in medical sectors need to predict the disease affected or not by finding accuracy calculation using machine learning techniques. The aim is to research machine learning based techniques for Parkinson disease by prediction leads to best accuracy with finding classification report. The analysis of dataset by supervised machine learning technique(SMLT) to capture several information's like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments and analyze the data validation, data cleaning/preparing and data visualization are going to be done on the whole given dataset. To propose, a machine learning-based method to accurately predict the disease by speech symptom by prediction results in the form of best accuracy and additionally compare the performance of various machine learning algorithms from the given hospital dataset with evaluation classification report, identify the result shows that GUI with best accuracy with precision, Recall ,F1 Score specificity and sensitivity.*

1.INTRODUCTION

Machine learning is to predict the future from past data. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python. It feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. Machine learning can be roughly separated in to three categories. There are supervised learning, unsupervised learning and reinforcement learning. Supervised learning program is both given the input data and the corresponding labeling to learn data has to be labeled by a human being beforehand. Unsupervised learning is no labels. It provided to the learning algorithm. This algorithm has to figure out the clustering of the input data. Finally, Reinforcement learning dynamically interacts with its environment and it receives positive or negative feedback to improve its performance.Data scientists use many different kinds of machine learning algorithms to discover patterns in python that lead to actionable insights. Classification predictive modeling is the task of approximating a mapping function from input variables(X) to discrete output variables(y). This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too.

1.1 What is Supervised Machine Learning?

Supervised Machine Learning is that the majority of practical machine learning uses supervised learning. Supervised learning is where have input variables (X) and an output variable (y) and use an algorithm to find out the mapping function from the input to the output is y = f(X). The goal is to approximate the mapping function so well that once you have new input file &#40;X&#41; that you simply can predict the output variables (y) for that data. Techniques of Supervised Machine Learning algorithms include logistic regression, multi-class classification, Decision Trees and support vector machines etc. Supervised learning requires that the info wont to train the algorithm is already labeled with correct answers. Supervised learning problems are often further grouped into Classification problems. This problem has as goal the construction of a succinct model that can predict the value of the dependent attribute from the attribute variables. The difference between the two tasks is the fact that the dependent attribute is numerical for categorical for classification. Given one or more inputs a classification model will attempt to predict the worth of 1 or more outcomes. A classification problem is when the output variable may be a category, like “red” or “blue”.

1.2 What is Parkinson's disease

Parkinson's disease a long term degenerative disorder of the central nervous system that affects the motor control of a patient by affecting predominately dopamine producing neurons in a specific area of the brain. The main problem in detecting the disease timely is the visible symptoms appear mostly at the later stage where cure no longer becomes possible. There is no correct reason proved yet that results to cause of Parkinson's, hence scientists are still conducting extensive research to find out its exact cause. Though some abnormal genes that become prominent due to elderly appear to lead to Parkinson's in some people but there is no evidence to proof this. Though there are a couple of procedures early Parkinson's detection. Dopamine transporter single-photon emission computed tomography can be used to effectively diagnosis Parkinson's by detecting amount of dopamine deficiency in the concerned patient's brain cell at a considerable early stage. Parkinson's disease (PD) affects approximately one million Americans and can cause several motor and non-motor symptoms. One of the secondary motor symptoms that people with PD may experience is change in speech, or speech difficulty. Not everyone with PD experiences same symptoms, and not all patients will have changes in their speech.

2. MODULES DESCRIPTION

2.1 VARIABLE IDENTIFICATION PROCESS AND DATA VALIDATION PROCESS

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the info volume is large enough to be representative of the population, you'll not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of knowledge wont to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters. The validation set is employed to guage a given model, but this is often for frequent evaluation. It as machine learning engineers uses this data to fine-tune the model hyper parameters. Data collection, data analysis, and therefore the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the method of knowledge identification, it helps to know your data and its properties; this data will assist you choose which algorithm to use to create your model. For example, time series data can be analyzed by regression algorithms; classification algorithms can be used to analyze discrete data. (For example to show the data type format of given dataset)

2.1.1 DATA VALIDATION/CLEANING/PREPARING PROCESS

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to form the simplest use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process. The primary goal of knowledge cleaning is to detect and take away errors and anomalies to extend the worth of knowledge in analytics and deciding .

2.1.2 DATA PREPROCESSING

Data Preprocessing may be a technique that's wont to convert the data into a clean data set. In other words, whenever the info is gathered from different sources it's collected in raw format which isn't feasible for the analysis. To achieving better results from the applied model in Machine Learning method of the data has to be in a proper manner. Some specified Machine Learning model needs information during a specified format; for instance , Random Forest algorithm doesn't support null values. Therefore, to execute random forest algorithm null values need to be managed from the first data set. And another aspect is that data set should be formatted in such how that quite one Machine Learning and Deep Learning algorithms are executed in given dataset.

2.2 DATA VISUALIZATION

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed specialise in quantitative descriptions and estimations of knowledge . Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and going to know a dataset and may help with identifying patterns, corrupt data, outliers, and far more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

2.3 LOGISTIC REGRESSION

It is a statistical procedure for analyzing a knowledge set during which there are one or more independent variables that determine an outcome. The goal of logistic regression is to seek out the simplest fitting model to explain the connection between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a group of independent (predictor or explanatory) variables. Logistic regression may be a Machine Learning classification algorithm that's wont to predict the probability of a categorical variable . In logistic regression, the variable may be a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).

2.3.1 DECISION TREE

Decision tree builds classification or regression models within the sort of a tree structure. It breaks down a knowledge set into smaller and smaller subsets while at an equivalent time an associated decision tree is incrementally developed. Decision tree builds classification or regression models within the sort of a tree structure. Each time a rule is learned, the tuples covered by the principles are removed.This process is sustained on the training set until meeting a termination condition. It is constructed in a top-down recursive divide-and-conquer manner. All the attributes should be categorical. Otherwise, they should be discretized in advance. Attributes within the top of the tree have more impact towards within the classification and that they are identified using the knowledge gain concept. A decision tree are often easily over-fitted generating too many branches and should reflect anomalies thanks to noise or outliers.

2.4 SUPPORT VECTOR MACHINES

A classifier that categorizes the info set by setting an optimal hyper plane between data. I chose this classifier because it is incredibly versatile within the number of various kernelling functions which will be applied and this model can yield a high predictability rate. Support Vector Machines are perhaps one among the foremost popular and talked about machine learning algorithms. They were extremely popular round the time they were developed within the 1990s and still be the go-to method for a high-performing algorithm with little tuning.

2.4.1 RANDOM FOREST

Random forest may be a sort of supervised machine learning algorithm supported ensemble learning. Ensemble learning may be a sort of learning where you join differing types of algorithms or same algorithm multiple times to make a more powerful prediction model. The random forest algorithm combines multiple algorithm of an equivalent type i.e. multiple decision trees, leading to a forest of trees, hence the name "Random Forest". The random forest algorithm are often used for both regression and classification tasks.

2.5 K-NEAREST NEIGHBOR

K-Nearest Neighbor is a supervised machine learning algorithm which stores all instances correspond to training data points in n-dimensional space. The k-nearest-neighbors algorithm may be a classification algorithm, and it's supervised: it takes a bunch of labeled points and uses them to find out the way to label other points. To label a replacement point, it's at the labeled points closest thereto new point (those are its nearest neighbors), and has those neighbors vote, so whichever label the most of the neighbors have is that the label for the new point (the “k” is that the number of neighbors it checks). Makes predictions about the validation set using the entire training set. KNN makes a prediction about a new instance by searching through the entire set to find the k “closest” instances.

2.5.1 NAIVE BAYES ALGORITHM

The Naive Bayes algorithm is an intuitive method that uses the possibilities of every attribute belonging to every class to form a prediction. It is the supervised learning approach you'd come up with if you wanted to model a predictive modeling problem probabilistically. This is a strong assumption but results in a fast and effective method. The probability of a category value given a worth of an attribute is named the contingent probability . By multiplying the conditional probabilities together for every attribute for a given class value, we've a probability of a knowledge instance belonging thereto class. To make a prediction we will calculate probabilities of the instance belonging to every class and choose the category value with the very best probability.Naive Bayes may be a statistical classification technique supported Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is that the fast, accurate and reliable algorithm. Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features. For example, a loan applicant is desirable or not depending on his/her income, previous loan and transaction history, age, and location. Even if these features are interdependent, these features are still considered independently. This assumption simplifies computation, and that's why it is considered as naive. This assumption is called class conditional independence.

2.6 GRAPHICAL USER INTERFACE

Tkinter is a python library for developing GUI (Graphical User Interfaces). We use the tkinter library for creating an application of UI (User Interface), to create windows and all other graphical user interface and Tkinter will come with Python as a standard package, it can be used for security purpose of each users or accountants. There will be two kinds of pages like registration user purpose and login entry purpose of users.It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare.

2.6.1 ACCURACY CACULATION

Table-1:Performance measurements of ML algorithm for speech

Parameters	LR	DT
Precision	0.69	0.76
Recall	0.73	0.87
F1-Score	0.71	0.81
Sensitivity	0.73	0.866
Specificity	0.88	0.90
Accuracy (%)	84.74	89.83

Table-2: Performance measurements of ML algorithm for tremor

Parameters	LR	DT
Precision	1	1
Recall	0.80	1
F1-Score	0.89	1
Sensitivity	0.8	1
Specificity	1	1
Accuracy (%)	95.83	100

Table-3: Performance measurements confusion matrix for speech

Parameters	LR	DT
TP	39	40
TN	11	13
FP	4	2
FN	5	4
TPR	0.88	0.90
TNR	0.73	0.86
FPR	0.26	0.13
FNR	0.11	0.09
PPV	0.90	0.95
NPV	0.98	0.76

Table-4: Performance measurements confusion matrix for tremor

Parameters	LR	DT
TP	19	19
TN	4	5
FP	1	0
FN	0	0
TPR	1	1
TNR	0.8	1
FPR	0.2	0
FNR	0	0
PPV	0.95	1
NPV	1	1

3. SYSTEM TECHNIQUES

Design is meaningful engineering representation of something that's to be built. Software design is a process design is the perfect way to accurately translate requirements in to a finished software product. Design creates a representation or model, provides detail about software arrangement, architecture, interfaces and components that are necessary to implement a system.

SYSTEM ARCHITECTURE

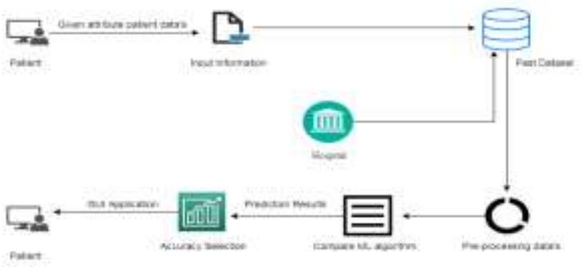

Fig-1:Architecture Diagram

HARDWARE REQUIREMENTS

➢ **Processor**	: i3/i4
➢ **Hard disk**	: minimum 300 GB
➢ **RAM**	: minimum 4 GB

SOFTWARE REQUIREMENTS

Operating System	: Windows
Tool Notebook	: Anaconda with Jupyter

SNAPSHOT

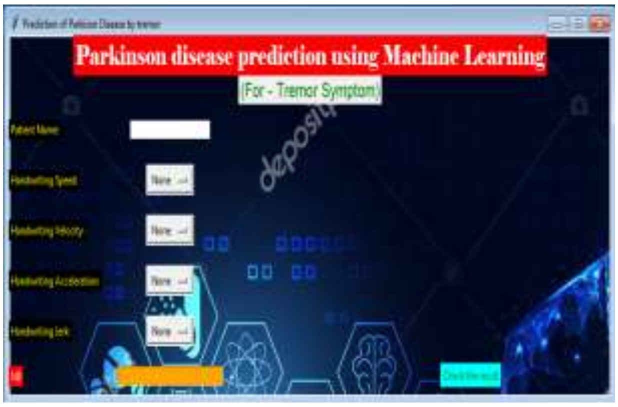

Fig-2:Parkinson disease prediction for tremor symptom

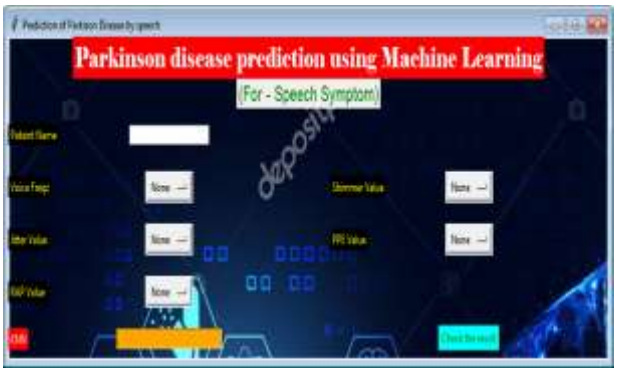

Fig-3: Parkinson disease prediction for speech symptom

ADVANTAGES

The scope of this project is to investigate a dataset of Parkinson normal person records and patient records for medical field using machine learning technique. To analyzing the prediction of Parkinson disease person is more accuracy with comparing algorithm and try to reduce the doctor's risk of factor behind detecting the patient. Easy to predicting the diagnose Parkinson with doctors can detecting the patient testing result time is reduced

APPLICATION

A hospital wants to automate the diagnosis of patient (real time) based on the patient detail provided while filling online/ offline application form. To automate this process, they have given a problem to identify the patient segments, those are have detecting disease to list to show doctors.

FUTURE ENHANCEMENT

Hospitals want to automate the detecting the disease persons from eligibility process (real time) based on the account detail.To automate this process by show the prediction result in web application or desktop application.To optimize the work to implement in Artificial Intelligence environment.

CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score will be finding out. This brings some of the following insights about diagnose the Parkinson disease. Early diagnosis of Parkinson's is most important for the patient to reduce its impact. It presented a prediction model with the aid of artificial intelligence to improve over human accuracy and provide with the scope of early detection. It can be inferred from this model that, area analysis and use of machine learning technique is useful in developing prediction models that can help a doctor reduce the long process of diagnosis and eradicate any human error.

REFERENCES

[1] X. Feng et al., “A language-independent neural network for event detection,” Sci. China Inf. Sci., vol. 61, no. 9, pp. 92–106, 2018

[2] F. Karim et al., “LSTM fully convolutional networks for time series classification,” IEEE Access, vol. 6, pp. 1662–1669, 2018.

[3] W. Liu et al., “Learning efficient spatial-temporal gait features with deep learning for human identification,” Neuroinformatics, vol. 16, pp. 457– 471, 2018.

[4] S. Yeung et al., “Every moment counts: Dense detailed labeling of actions in complex videos,” Int. J. Comput. Vis., vol. 126, no. 2–4, pp. 375–389, 2018.

[5] M. M. Hassan et al., “A robust human activity recognition system using smartphone sensors and deep learning,” Future Gener. Comput. Syst., vol. 81, pp. 307–313, 2018.

.