Tech Saksham

Final Project Report

Full Stack Web Development

“Multiple Disease Prediction using ML(Python)”

“IIIT RGUKT RK VALLEY”

ROLL NO	NAME
R170413	T .Dharani

Poovaragavan Velumani

Master Trainer

ABSTRACT

Many of the existing machine learning models for health care analysis are concentrating on one disease per analysis. Like one analysis if for diabetes analysis, one for brain diseases like that. There is no common system where one analysis can perform more than one disease prediction. In this project proposing a system which used to predict multiple diseases by using Streamlit in Python. In this article used to analyze Diabetes analysis, Parkinson disease analysis. To implement multiple disease analysis used machine learning algorithms, Streamlit and web designing. Python pickling is used to save the model behaviour and python unpickling is used to load the pickle file whenever required. The importance of this project analysis in while analyzing the diseases all the parameters which causes the disease is included so it possible to detect the maximum effects which the disease will cause. For example for diabetes analysis in many existing systems considered few parameters like age, sex, BMI , insulin, glucose, blood pressure, diabetes pedigree function, pregnancies, considered in addition to age, sex, BMI, insulin, glucose, blood pressure, diabetes pedigree function, pregnancies included serum creatinine, potassium, Glasgow Com Scale, heart rate/pulse Rate, respiration rate, body temperature, low density lipoprotein (LDL), high density lipoprotein (HDL), TG (Triglycerides).Final models behaviour will be saved as python pickle file. After user accessing this website, the user has to send the parameters of the disease along with disease name. This website will invoke the corresponding model and returns the status of the patient. The importance of this analysis to analyze the two diseases, so that to monitor the patient’s condition and warn the patients in advance to decrease mortality ratio

INDEX

CHAPTER 1

INTRODUCTION

1.1 Overview

During a lot of analysis over existing systems in healthcare analysis considered only one disease at a time.When any organization wants to analyse their patient's health reports then they have to deploy many models. The approach in the existing system is useful to analyse only particular disease. Now a day's mortality got increased due to exactly not identifying exact disease. Even the patient got cured from one disease may be suffering from another disease. Some existing systems used few parameters while analysing the disease. Due to that may be not possible to identify the diseases which will be caused due to the effect of that disease. For example, due to diabetes, there may be chance of neuropath, retinopathy, hearing loss, and dementia. In this article considered Diabetes analysis and Parkinson disease data sets. In future many other diseases like skin diseases can be included, fever related diseases and many more. This analysis is flexible that later included many diseases for analysis. While adding any new disease analysis to this existing Website, the developer has to add the model file related to the analysis of the new disease. When developing new disease the developer have to prepare python picking to save model behaviour. When using this Streamlit in Python, the developer can load pickled file to retrieve the model behaviour. When user wants to analyse the patient's health condition either then can predict a particular disease or if the report contains parameters which are used to predict other diseases then this analysis will produce maximum identification of relevant diseases. The aim of this article is used to prevent mortality ratio increasing day by day by warning the patients in advance based on their health conditions. Due to many diseases models and predictions done at one place cost of patient analysis can be reduced.

1.2 Feature

Machine Learning may be a sub-area of AI, whereby the term refers to the power of IT systems to independently find solutions to problems by recognizing patterns in databases. In other words: Machine Learning enables IT systems to acknowledge patterns in the idea of existing algorithms and data sets and to develop adequate solution concepts. Therefore, in Machine Learning, artificial knowledge is generated on the idea of experience. In order to enable the software to independently generate solutions, the prior action of people is important. For example, the required algorithms and data must be fed into the systems in advance and the respective analysis rules for the recognition of patterns in the data stock 6 must be defined. Once these two steps have been completed, the system can perform the following tasks by Machine Learning:

1.Finding, extracting and summarizing relevant data.

2.Making predictions based on the analysis data.

3.Calculating probabilities for specific results.

In the course of monitored learning, example models are defined beforehand.So as to make sure an adequate allocation of the knowledge to the respective model groups of the algorithms, these then need to be specified. In other words, the system learns on the idea of given input and output pairs. Within the course of monitored learning, a programmer, who acts as a sort of teacher, provides the acceptable values for specific input. The aim is to coach the system within the context of successive calculations with different inputs and outputs to determine connections.Supervised learning is where you've got input variables (X) and an output variable (Y) and you employ an algorithm to find out the mapping function from the input to the output. Y = f(X) The goal is to approximate the mapping function so well that once you have a new input file (X) that you simply can predict the output variables (Y) for that data. It's called supervised learning because the

method of an algorithm learning from the training dataset is often thought of as an educator supervising the training process. We all know the correct answers, the 7 algorithm iteratively makes predictions on the training data and is corrected. Learning stops when the algorithm achieves a suitable level of performance. Techniques of Supervised Machine Learning algorithms include linear and logistic regression, multi-class classification, Decision Tree, and Support Vector Machine. Supervised Learning problems are a kind of machine learning technique often further grouped into Regression and Classification problems. The difference between these two is that the dependent attribute is numerical for regression and categorical for classification:

● Regression Linear regression could also be a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and thus the only output variable (y). More specifically, that y is usually calculated from a linear combination of the input variables (x).When there's one input variable (x), the tactic is mentioned as simple linear regression. When there are multiple input variables, literature from statistics often refers to the tactic as multiple linear regression.

● Classification Classification could also be a process of categorizing a given set of data into classes, It is often performed on both structured or unstructured data. the tactic starts with predicting the category of given data points. The classes are often mentioned as target, label, or categories. In short, classification either predicts categorical class labels or classification data supported the training set and thus the values(class labels) in classifying attributes and uses it in classifying new data. There is a variety of classification models. Classification models include Logistic Regression, Decision Tree, Random Forest, Gradient Boosted Tree, One vs.-One, and Naïve Bayes,SVM.

In unsupervised learning, AI learns without predefined target values and without rewards. It's mainly used for learning segmentation (clustering). The

machine tries to structure and type the info entered consistent with certain characteristics. For instance, a machine could (very simply) learn that coins of various colors are often sorted consistent with the characteristic "color" so as to structure them. Unsupervised Machine Learning algorithms are used when the knowledge used to train is neither classified nor labeled. The system doesn't 8 determine the right output but it explores the data and should draw inferences from datasets to elucidate hidden structures from unlabeled data. Unsupervised Learning is that the training of Machines using information that's neither classified nor labeled and allowing the algorithm to act thereon information without guidance. Unsupervised Learning is accessed into two categories of algorithms inherent grouping in the data such as grouping customers by purchasing behaviour.

● Clustering A clustering problem is where you would like to get the inherent grouping in the data such as grouping customers by purchasing behaviour.

● Association An Association rule learning problem is where you would wish to get rules that describe large portions of your data such as folks that buy X also tend to shop for Y.

1.3 Scope

Diabetes is a very serious disease with many life threatening consequences, but if it is taken care of properly, diabetes can live a normal life. Parkinson’s disease is the second most dangerous neurodegenerative disease which has no cure till now and to make it reduce prediction is important. In this project, we have used prediction model(SVM) to predict the Diabetes, Parkinson’s disease which are Machine Learning Technique. The dataset is trained using these models and we also compared these different models built. We have used the dataset that contains some features of the patients which is available in the Kaggle website. The dataset consists of more than 700 features and 750 patient details. The models are built using the some best features which were identified by feature selection. This

system we designed can make the predictions of the Diabetes and Parkinson's disease.

1.4 Future Work

In future, these models can be trained with different datasets that have best features and can be predicted more accurately. If the accuracy rate increases, it can be used by the laboratories and hospitals so that it is easy to predict in early stages. This models can be also used with different medical and disease datasets. In future the work can be extended by building a hybrid model that can find more than three diseases with an accurate dataset and that dataset has common features of two diseases. In future the work can extended to build a model that may extract more important features among all features in the dataset so that it produce more accuracy.

CHAPTER 2

SERVICES AND TOOLS REQUIRED

2.1 Services Used

Introduction to Python:

Python is an interpreter, high-level, general-purpose programming language. Python is simple and easy to read syntax emphasizes readability and thus reduces system maintenance costs. Python supports modules and packages, which promote system layout and code reuse. It saves space but it takes a rather higher time when its code is compiled. Indentation must be taken care while coding. Python does the following:

➔ Python are often used on a server to make web applications.

➔ It connects the database systems. It also read and modify files.

➔ It often able to handle big data and perform complex mathematics.

➔ It can be used for production-ready software development. Python has many inbuilt library functions that can be used easily for working with machine learning algorithms. All the necessary python libraries must be pre-installed using “pip” command.

Introduction to Streamlit:

Streamlit is meant to help Data scientists or machine learning engineers who are not web developers and they're not interested in spending weeks learning to use these frameworks to build web apps. The best thing about Streamlit is that you don't even need to know the basics of web development to get started or to create your first web application. So 17 if you're somebody who's into data science and you want to deploy your models easily, quickly, and with only a few lines of code, Streamlit is a good fit

2.2 Tools and Softwares used

1. Software: Spyder Google Colab

2. Tools : Web Browser ,Google Chrome

3.Python Libraries : numpy, pandas, streamlit , sklearn , pickle.

2.2.1 Numpy

NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the elemental package for scientific computing with Python. It contains various features including these important ones:

• A powerful N-dimensional array object

• Sophisticated (broadcasting) functions

• Tools for integrating C/C++ and Fortran code

• Useful linear algebra, Fourier transform, and random number capabilities Besides its obvious scientific uses, NumPy also can be used as an efficient multidimensional container of generic data

2.2.2 Pandas

Pandas is an open-source library that's built on top of NumPy library. It is a Python package that gives various data structures and operations for manipulating numerical data and time series. It is fast and it has high-performance & productivity for users. It provides high-performance and is easy-to-use data structures and data analysis tools for the Python language. Pandas is employed during a wide range of fields including academic and commercial domains including economics, Statistics, analytics, etc.

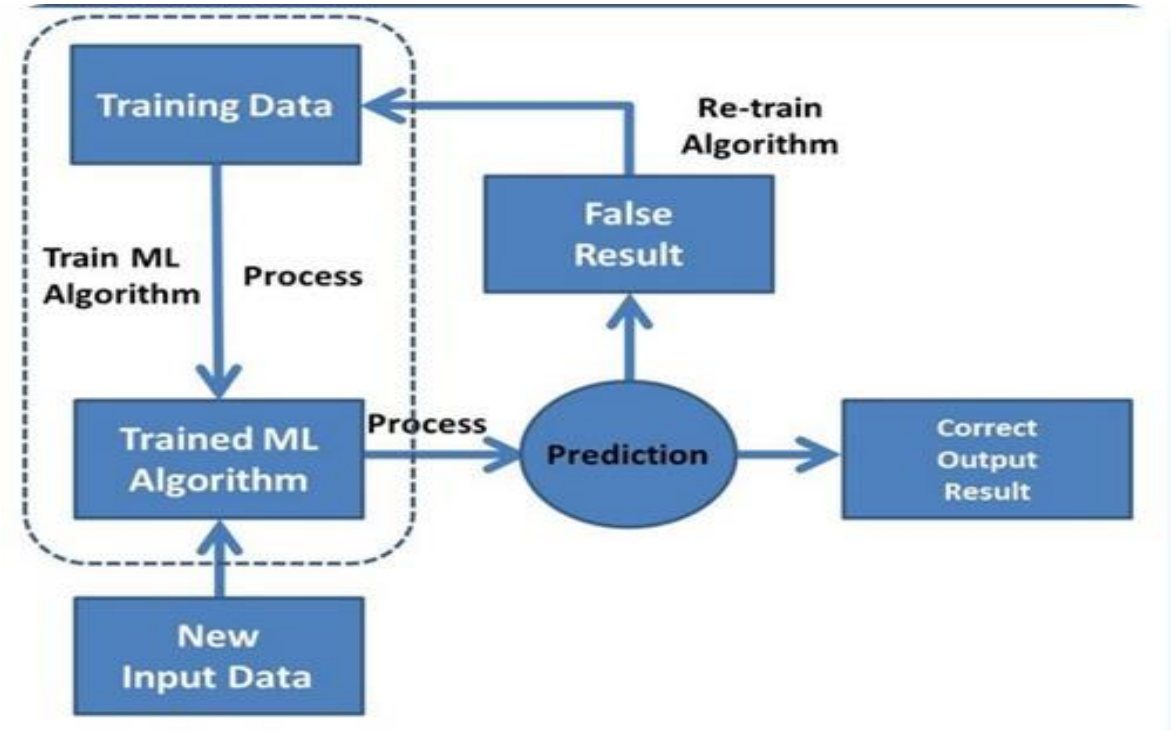
2.2.3 Sklearn

Scikit-learn (Sklearn) is that the most useful and robust library for machine learning in Python. It is an open-source Python library that implements a variety of machine learning, pre-processing, cross-validation and visualization algorithms employing a unified interface. Sklearn provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via 36a consistence interface in Python. This library, which is essentially written in Python, is made upon NumPy, SciPy and Matplotlib

CHAPTER 3

PROJECT ARCHITECTURE

3.1 Architecture



Machine learning has given computer systems the ability to automatically learn without being explicitly programmed. In this project, we used machine learning algorithm (SVM). The architecture diagram describes the highlevel overview of major system components and important working relationships. It represents the flow of execution and it involves the following five major steps: The architecture diagram is defined with the flow of the process which is used to refine the raw data and used for predicting the data. The next step is preprocessing the collected raw data into an understandable format. Then we have to train the data by splitting the dataset into train data and test data. The Diabetes and Parkinson's data is evaluated with the application of a machine

learning algorithm that is SVM and the classification accuracy of this model is found. After training the data with these algorithms we have to test on the same algorithms. Finally, the result of these algorithm is compared on the basis of classification accuracy

. • Speech Dataset

• Pre-processing data

• Training data

• Apply Machine Learning Algorithm(SVM)

• Testing Data

CHAPTER 4

ARCHITECTURE BLOCKS DETAIL WORKING

4.1 Blocks

Speech Dataset

The main aim of this step is to spot and acquire all data-related problems. During this step, we'd like to spot the various data sources, as data are often collected from various sources like files and databases. The number and quality of the collected data will determine the efficiency of the output. The more are going to be the info, the more accurate are going to be the prediction. We've collected our data from the Kaggle website.

we can see the speech dataset that has collected from kaggle website. This acquired dataset has around 755 patient's data and each row has 755 different features. But in this paper, we chosen 10 main features that required to find the prediction. Reading the dataset from the CSV file into notebook The dataset we chose is in the form of CSV (Comma Separated Value) file. After acquiring the data our next step is to read the data from the CSV file into the Google colab also called a Python notebook. Python notebook is used in our project for data pre-processing, features selection, and for model comparison.

Data Pre-Processing

The main aim of this step is to study and understand the nature of data that was acquired in the previous step and also to know the quality of data. A realworld data generally contains noises, missing values, and maybe in an unusable format that cannot be directly used for machine learning models. Data preprocessing is a required task for cleaning the data and making it suitable for a machine learning

model which also increases the accuracy and efficiency of a machine learning model. Identifying duplicates in the dataset and removing them is also done in this step. Actually, in this dataset, we have 755 features out of which some may not be useful in building our model. So, we have to leave out all those unnecessary features which are not responsible to produce the output. If we take more features in this model the accuracy we got is less. When we check the correlation of the features, some of them are the same.

Training data

Splitting the dataset into Training set and testing set: In machine learning data pre-processing, we have to break our dataset into both training set and test set. This is often one among the crucial steps of knowledge pre-processing as by doing this, we will enhance the performance of our machine learning model. Suppose, if we've given training to our machine learning model by a dataset and that we test it by a totally different dataset. Then, it'll create difficulties for our model to know the correlations between the models. If we train our model alright and its training accuracy is additionally very high, but we offer a replacement dataset there to, then it'll decrease the performance. So we always attempt to make a machine learning model which performs well with the training set and also with the test dataset.

Apply Machine Learning Algorithm

Now, we've both the train and test data. The subsequent step is to spot the possible training methods and train our models. As this is often a classification problem, we've used classification method SVM. Each algorithm has been run over the training dataset and their performance in terms of accuracy is evaluated alongside the prediction wiped out the testing data set.

Support vector machine

SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression 14 problems. However, primarily, it is used for Classification problems in Machine Learning. 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 we 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 support vectors, and hence the algorithm is termed as Support Vector Machine. Support Vector Machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. SVMs have their unique way of implementation as compared to other machine learning algorithms. Lately, they are extremely popular because of their ability to handle multiple continuous and categorical variables.

Testing Data

Once multiple disease Prediction model has been trained on the preprocessed dataset, then the model is tested using different data points. In this testing step, the model is checked for correctness and accuracy by providing a test dataset to it. All the training methods need to be verified for finding out the best model to be used test dataset. These predicted values on testing data are used for model comparison and accurate calculation.

CONCLUSION

Diabetes is a very serious disease with many life threatening consequences, but if it is taken care of properly, diabetes can live a normal life. Parkinson's disease is the second most dangerous neurodegenerative disease which has no cure till now and to make it reduce prediction is important. In this project, we have used prediction model(SVM) to predict the Diabetes, Parkinson's disease which are Machine Learning Technique.

The dataset is trained using these models and we also compared these different models built. We have used the dataset that contains some features of the patients which is available in the Kaggle website. The dataset consists of more than 700 features and 750 patient details. The models are built using the some best features which were identified by feature selection. This system we designed can make the predictions of the Diabetes and Parkinson's disease.

REFERENCES

[1] Md. Redone Hassan,etal, “A Knowledge Base Data Mining based on Parkinson’sDisease” International Conference on System Modelling & Advancement in ResearchTrends, 2019.

[2] R. P. Duncan, A. L. Leddy, J. T. Cavanaugh et al., “Detecting and predicting balance decline in Parkinson disease: a prospective cohort study” Journal of Parkinson’s Disease,vol-5, no. 1, pp. 131–139, 2015.

[3] T. J. Wroge, Y. Özkanca, C. Demiroglu, D. Si, D. C. Atkins and R. H. Ghomi,"Parkinson’s Disease Diagnosis Using Machine Learning and Voice," IEEE Signal Processing in Medicine and Biology Symposium (SPMB), pp.1-7, doi:10.1109/SPMB.2018.8615607, 2018.

[4] Yashoda and M. Kannan, "Analysis of a Population of Diabetic Patients Databases in Waikato", International Journal of Scientific & Engineering Research, vol. 2, no. 5, 2011.

[5] A. Ayer, J. S and R. Sumbala, "Diagnosis of Diabetes Using Classification Mining Techniques", IJDKP, vol. 5, no. 1, pp. 01-14, 2015.

[6] NIyati Gupta, A. Rawal, and V. Narasimhan, “Accuracy, Sensitivity andSpecificity Measurement of Various Classification Techniques on Healthcare Data", IOSR Journal of Computer Engineering, vol. 11, no. 5, pp. 70-73,2013

CODE

https://github.com/Dharani170413/code.git