DIABETES DISEASE PREDICTION

Project Report

Submitted in the partial fulfillment of the requirements for the

award of the degree of

Bachelor of Technology

in

COMPUTER SCIENCE AND ENGINEERING

180030732 SHAIK KHATUNA

by

under the supervision of

Dr.P Ithaya Rani

K L E F

KONERU LAKSHMAIAH EDUCATION FOUNDATION

(Deemed to be university estd, u/s, 3 of the UGC Act, 1956)

(NAAC Accredited "A" Grade University)

Green Fields, Vaddeswaram- 522502, Guntur(Dist), Andhra Pradesh, India.

Nov,2021

KONERU LAKSHMAIAH EDUCATION FOUNDATION

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Declaration

The Project Report entitled "**DIABETES DISEASE PREDICTION** "is a record of bonafide work of **180030732-SK KHATUNA**, submitted in partial fulfillment for the award of B.Tech in **COMPUTER SCIENCE ENGINEERING** to the K L University. The results embodied in this report have not been copied from any other departments/University/Institute.

180030732 SK KHATUNA

KONERU LAKSHMAIAH EDUCATION FOUNDATION

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Certificate

This is to certify that the Project Report entitled "**DIABETES DISEASE PREDICTION** " is being submitted by **180030732-SK KHATUNA** submitted in partial fulfillment for the award of B.Tech in **COMPUTER SCIENCE ENGINEERING** to the K L University is a record of bonafide work carried out under our guidance and supervision.

The results embodied in this report have not been copied from any other departments/ University/Institute.

Mr. Hari Kiran Vege
Head of Department

Dr.P Ithaya Rani
Project Supervisor

ACKNOWLEDGEMENT

We We desire to explicit our deepest sense of gratitude towards our President, **Sri Koneru Satyanarayana** for giving us an possibility to finish the path of work on the outset.

We We express our deep sense of gratitude to the loved Vice Chancellor **Dr. G. Pardha Saradhi Varma**, KLEF, for giving us the opportunity to complete our direction of labor.

We explicit a wholehearted gratitude to **Dr.k.Subbarao**, most important KLEF, for presenting us the conductive environment for carrying through our educational schedules and projects

We We specific our gratitude to **Mr.V.HariKiran**, Head of
the branch for computer technological know how and Engineering for presenting us with adequate centers, methods and method via which we are in a position to finish the work.

We are thankful and explicit our gratitude towards our guide of computer technological know-how and Engineering, for giving us an opportunity to paintings on the topic that we are able to accomplish.

We're grateful to the staff of **KLEF** for providing facilities to finish the assignment.

We acknowledge the assist of all those who have given encouragement and accomplice themselves in one way or the other inside the final touch of the dissertation work.

.

ABSTRACT

The The objective of Diabetes ailment Prediction mission is to diagnostically are expecting whether or not a affected person has diabetes or no longer, based totally on positive diagnostic measurements protected within the records set. several constraints have been located on the choice of these instances from a larger database. Diabetes has a exquisite deal of interest in scientific research. The analysis of Diabetes sickness is a difficult assignment, that may offer automated prediction about the Diabetes circumstance of patient in order that similarly remedy can be made powerful. especially most of the women develop gestational diabetes all through their pregnancy period. ladies with gestational diabetes don't usually have symptoms or might also chalk them up to pregnancy. most find out that they have it at some stage in a ordinary screening

INDEX

List of Figures

1. INTRODUCTION

1.1 Introduction

Machine studying is one of the packages of synthetic intelligence (AI) that provides computer systems, the capacity to learn mechanically and improve from enjoy in place of explicitly programmed. It specializes in developing computer packages that can get entry to records and use it to learn from themselves. the principle intention is to permit computer systems to study automatically without human intervention and also modify movements for that reason

Diabetes has a deal of attention in medical research. The diagnosis of Diabetes disease is a challenging task, which can offer automated prediction about the Diabetes condition of patient so that further treatment can be made effective. In particular most of the women develop gestational diabetes during their pregnancy period. Women with gestational diabetes don't usually have symptoms or may chalk them up to pregnancy. Most find out that they have it during a routine screening.

Gestational Diabetes is a situation wherein your blood sugar degrees come to be excessive at some point of being pregnant. It impacts up to ten% of ladies who are pregnant within the U.S. every year. There are two classes of gestational diabetes. girls with elegance A1 can control it via eating regimen and exercise. the ones who've elegance A2 want to take insulin or different medications. Gestational diabetes goes away when you deliver start. but it is able to have an effect on your child's fitness, and it increases your hazard of getting kind 2 diabetes later in lifestyles.

Type algorithms are very vital class of supervised device getting to know algorithms. these algorithms require a completely huge education set. those education data sets are which includes many capabilities or attributes which describe the person pattern. seeing that we are doing supervised getting to know algorithm. all of the training set are categorised successfully. The type algorithms including decision Tree Classifier, Gaussian Naive Bayes, Random woodland Classifier, k-Nearest Neighbor, k-way and Logistic Regression and many others., expand model with those statistics with many one-of-a-kind parameters. whilst we have a brand new unlabeled sample, we can use the version to expect the label of the new sample. those techniques are used to predict

2. SYSTEM ANALYSIS

2.1 Existing System

Hospitals maintain all the patient records. Even though, those records are not used in an efficient manner for diagnosis. To maintain the records in an efficient error free manner, the new proposed system is introduced**.**

2.1.1 Disadvantages:

1. Doesn't generate accurate and efficient results
2. Computation time is very high
3. Difficulty in maintenance of patient records
4. Lacking of accuracy may result in lack of efficient further treatment

2.2 Proposed System

To develop a system which will help practitioners to predict diabetes based on some diagnostic measurements like number of pregnancies, age, gender, blood pressure and so on. So, there is a need for developing a decision system which will help practitioners to predict whether a patient has diabetes or not in an easier way, so that further treatment can be made effectively. This proposed system not only accurately predicts diabetes but also reduces time for prediction.

2.2.1 Advantages:

1. Generates accurate and efficient results
2. Computation time is greatly reduced
3. Easy maintenance of patient records
4. Reduces manual work
5. Efficient further treatment
6. Automated prediction

2.3. System Requirements

2.3.1 Hardware Requirements:

- System Type : Intel(R) Core™2 i7-5500U CPU @ 2.40GHz
- Cache memory : 4MB(Megabyte)
- RAM : 8 gigabyte (GB)

2.3.2 Software Requirements:

- Operating System : Windows 10 Home, 64 bit Operating System
- Coding Language : Python
- Python distribution : Anaconda, Visual StudioCode

3. LITERATURE SURVEY

3.1 Machine Learning

Machine gaining knowledge of is one of the packages of synthetic intelligence (AI) that provides computers, the potential to study mechanically and enhance from revel in rather than explicitly programmed. It specializes in growing pc packages that may get right of entry to records and use it to analyze from themselves. the main intention is to permit computers to study mechanically with out human intervention and additionally regulate actions consequently.

3.2 Some machine learning methods

Gadget studying algorithms are often categorized as supervised and unsupervised.

- **Supervised machine learning algorithms** can Supervised system getting to know algorithms can apply what has been found out inside the beyond to new facts the use of classified examples to expect future activities. starting from the analysis of a recognized education information set, the getting to know algorithm produces an inferred feature to make predictions about the output values. The machine is able to provide targets for any new enter after enough training. The learning algorithm can also evaluate its output with the perfect, meant output and discover errors in an effort to modify the model accordingly.

- **Unsupervised machine learning algorithms** are In contrast, unsup23ervised gadget mastering algorithms are used whilst the records used to educate is neither categorized nor labeled. Unsupervised mastering research how systems can infer a function to explain a hidden structure from unlabeled statistics. The system doesn't parent out the proper output, but it explores the records and might draw inferences from datasets to explain hidden structures from unlabeled information.

- **Semi-supervised machine learning algorithms** The structures that use this technique are capable of drastically improve learning accuracy. typically, semi-supervised learning is chosen while the obtained classified information calls

for professional and relevant as sets which will train it / analyze from it. otherwise, acquiring unlabeled statistics normally doesn't require additional resources.

- **Reinforcement machine learning algorithms** is a getting to know algorithms is a studying approach that interacts with its environment by means of producing actions

error seek and behind schedule reward are the most applicable characteristics of reinforcement getting to know. This method lets in machines and software marketers to routinely decide the ideal behaviour inside a selected context with a view to maximize its overall performance. simple praise comments is required for the agent to research which motion is fine. that is called the reinforcement sign.

3.3 Applications of machine learning

1. Virtual Personal Assistants
2. Predictions while Commuting
3. Videos Surveillance
4. Social Media Services
5. Email Spam and Malware Filtering
6. Online Customer Support
7. Search Engine Result Refining
8. Product Recommendations
9. Online Fraud Detection

3.4 Prevalence of Diabetes Diseases

- The number of variety of people with diabetes rose from 108 million in 1980 to 422 million in 2014.
- The global the worldwide prevalence of diabetes* among adults over 18 years of age rose from 4.7% in 1980 to eight.5% in 2014.
- Diabetes is a chief reason of blindness, kidney failure, coronary heart attacks, stroke and lower limb amputation.
- In 2016, In 2016, an expected 1.6 million deaths had been at once as a result of diabetes. any other 2.2 million deaths had been as a consequence of excessive blood glucose in 2012.
- Almost half nearly 1/2 of all deaths attributable to excessive blood glucose arise earlier than the age of 70 years. WHO estimates that diabetes change.

3.5 Importance of machine learning in healthcare

The importance of gadget getting to know in health care is growing because of its capacity to technique big datasets efficaciously beyond the variety of human functionality,and then dependably convert analysis of that data into clinical insights that assist physicians in planning and providing care, which ultimately leads to better outcomes, reduces the charges of care, and increases sufferers pleasure. Using these types of advanced analysis, we can provide better information to doctors at the point of patient care.

3.6 Implementation of machine learning using Python

Python is a popular programming language. It was created in 1991 by Guido van Rossum.

It is used for:

- web development (server-side),
- software development,
- mathematics,
- system scripting.

The maximum latest fundamental model of Python is Python 3. but, Python 2, even though not being updated with whatever apart from safety updates, continues to be pretty popular.

It’s far possible to write Python in an incorporated improvement environment, along with Thonny, Pycharm, Netbeans or Eclipse, Anaconda that are in particular beneficial when dealing with large collections of Python documents.
Python Python changed into designed for its readability. Python uses new traces to complete a command, instead ofdifferent programming languages which often use semicolons or parentheses.

Python predicated on indentation, using white space, to define scope; including the scope of loops, functions and instructions. different programming languages frequently use curly-brackets for this motive.

In the older days, human beings used to perform gadget gaining knowledge of duties manually by coding all the algorithms and mathematical and statistical method. This made the technique time eating, tedious and inefficient. however inside the present day days, it is turn out to be very a great deal clean and efficientcompared to the olden days via numerous python libraries, frameworks, and modules. nowadays, Python is one of the maximum popular programming languages for this assignment and it has replaced many languages in the enterprise, one of the motive is its big series of libraries
. Python libraries that used in Machine Learning are:

- Numpy
- Scipy
- Scikit-learn
- Theano
- TensorFlow
- Keras
- PyTorch
- Pandas
- Matplotlib

NumPy is a very famous python library for big multi-dimensional array and matrix processing, with the helpof a large series of excessive-stage mathematical capabilities. it is very beneficial for fundamental scientific computations in machine mastering. it's far specifically beneficial for linear algebra, Fourier rework, and random variety abilities. high-give up libraries like TensorFlow uses NumPy internally for manipulation of Tensors

SciPy is a very completely famous library amongst device getting to know enthusiasts because it carries exceptional modules for optimization, linear algebra, integration and records. there may be a difference among the SciPy library and the SciPy stack. The SciPy is one of the center packages that make up the SciPy stack. SciPy is also very beneficial for image manipulation.

Skikit-learn is one of the most popular machine mastering libraries for classical system gaining knowledge of algorithms. it is built on pinnacle of two simple Python libraries, NumPy and Scikit-analyse supports maximum of the supervised and unsupervised getting to know algorithms. Scikit-learn can also be used for facts-mining and facts-analysis, which makes it a remarkable device who is beginning out with gadget studying.

Sci kit - analyze supports maximum of the supervised and unsupervised getting to know algorithms. Sci-kit-learn can also be used for facts-mining and facts-analysis, which makes it a remarkable device who is beginning out with gadget studying.

This is a famous python library that is used to define, evaluate and optimize mathematical expressions involving multidimensional arrays in an efficient way. it's miles executed by way of optimizing the usage of CPU and GPU. it's far substantially used for unit-trying out and self- verification to discover and diagnose exclusive forms of mistakes. This a totally effective library that has been used in large-scale computationally intensive clinical tasks for a long time but is straightforward and approachable enough to be utilized by individuals for his or her own projects.

Tensor Flow is the most useful and very popular open-source library for high overall performance numerical computation evolved by means of the Google mind group in Google. as the call suggests, Tensor flow is a framework that entails defining and strolling computations involving tensors. it may train and run deep neural networks that can be used to increase numerous AI packages. Tensor Flow is widely used inside the subject of deep studying research and application.

Keras is a completely famous system mastering library for Python. it's miles a high-stage neural networks API capable of running on pinnacle of Tensor Flow, CNTK, or The ano. it may run seamlessly on each CPU and GPU. Keras makes it without a doubt for ML novices to build and design a Neural network. one of the first- class issue approximately Keras is that it permits for easy and speedy prototyping.

Py Torch is a popular open-source system getting to know library for Python based on Torch, which is an open-supply gadget getting to know library that is carried out in C with a wrapper in Lua. It has an intensive choice of gear and libraries that helps on p.c vision, herbal Language Processing(NLP) and lots of extra ML packages. It allows developers to perform computations on Tensors with GPU acceleration and also helps in growing computational graphs.

Pandas is a famous Python library for information analysis. It isn't always directly associated with machine studying. As we realize that the information set need to be prepared before education. In this case, Pandas comes handy as it was advanced

excessive-stage facts structures and wide range equipment for statistics evaluation. It gives many in-built techniques for groping, combining and filtering information.

Mat plot lib is a very popular Python library for records visualization. Like Pandas, it isn't always directly related to system learning. It particularly is available in reachable while a programmer desires to visualize the patterns inside the facts. it's miles a 2d plotting library used for developing 2nd graphs and plots. A module named pyplot makes it smooth for programmers for plotting as it presents capabilities to control line patterns, font homes, formatting axes, etc. It offers diverse styles of graphs and plots for facts visualization, histogram, error charts, bar chats, and many others.

4. SYSTEM DESIGN

Scope of the project

The scope of this device is to maintain affected person information in data sets, teach the model the usage of the big amount of statistics found in data sets and expect whether or not presence or absence of sickness on new information all through trying out.

4.1 Scope of the project

09 attributes are used in dataset as follows:

1. Pregnancies

2. Glucose

3. Blood Pressure

4. Skin Thickness

5. Insulin

6. BMI: To find the Body Mass Index of a person

7. Diabetes Pedigree Function

8. Age

9. Outcome: If diabetic is present the out is 1 otherwise 0

The data set contains 769 instances

The data set is converted into a csv (comma separated values) report

.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigre	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0

Fig:4.2.1 Data set

Dataset Description

1) Pregnancies: Number of times pregnant
2) Glucose: Plasma glucose attentions a 2 hours in an oral glucose tolerance check
3) Blood Pressure: Diastolic blood pressure (mm Hg)
4) Skin Thickness: Triceps skin fold thickness (mm)
5) Insulin: 2-Hour serum insulin (mu U/ml)
6) BMI: Body mass index (weight in kg/(height in m)^2)
7) Diabetes Pedigree Function: Diabetes pedigree function
8) Age: Age(years)
9) Outcome: class variable (zero or 1) 268 of 768 are 1, the others are 0

4.2 Data Pre-processing

Records pre-processing refers to the changes applied to our statistics before feeding it to the set of rules. data Pre-processing is a technique this is used to transform the uncooked data into a easy records set. In different phrases, on every occasion the information is collected from one-of-a-kind sources it is collected in uncooked format which isn't feasible for the analysis. Pre-processing is step one whilst developing the system mastering version. it is the procedure of converting uncooked data set into cleaned data set. uncooked statistics contains noise, lacking values, duplicate values which isn't always appropriate for machine learning model. So, preprocessing is required for cleansing the facts and making it suitable for system gaining knowledge of version.

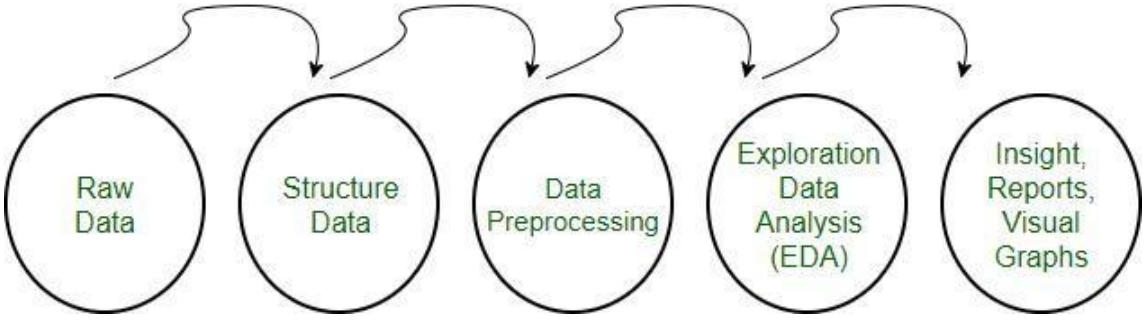


Fig:4.2.1 Data Pre-processing

Need of Data Pre-processing

For reaching higher effects from the implemented version in gadget learning projects the format of the information has to be in a proper manner. some targeted system mastering version needs facts in a detailed layout. as an example, Random wooded area set of rules does not help null values, consequently to execute random wooded area algorithm null values must be controlled from the origin

Another any other issue is that facts set must be formatted in any such manner that multiple device learning and Deep gaining knowledge of algorithms

.

4.3 Classification

It is a method of categorizing records into given classes. Its primary aim is to become aware of the magnificence of our new statistics.

4.3.1 Missing Values

Research research on information mining has led to the components of numerous records mining algorithms. These algorithms can be without delay used on a dataset for developing some models or to attract essential conclusions and inferences from that dataset. a few famous facts mining algorithms are choice tree, Naive Bayes, manner, synthetic neural community etc.

.

1. **Decision Tree**:

Decision selection Tree analysis is a popular, predictive modeling device that has packages spanning a number of specific regions.
In standard, decision bushes are constructed through an algorithmic method that identifies ways to split a information set based totally on extraordinary conditions. it's far one of the most broadly used and sensible strategies for supervised studying. decision trees are a non-parametric supervised getting to know approach used for each category and regression duties.

2. **Naive Bayes (NB):**

It is a simple technique for constructing classifiers. It is a probabilistic classifier based on Bayes' theorem. All Naive Bayes classifiers count on that the cost of any unique function is independent of the value of every other feature, given the elegance variable. Bayes theorem is given as follows: percent|X) = P(X|C)
* p.c)/P(X), wherein X is the data tuple and C is the magnificence such that P(X) is constant for all classes. although it assumes an unrealistic situation that characteristic values are conditionally independent, it play exceptionally properly on huge data sets in which this circumstance is assumed and holds.

3. Random Forest:

Random Forests are an ensemble studying technique (additionally concept of as a form of nearest neighbour predictor) for classification and regression techniques. It builds a couple of selection bushes and then merges them collectively in-order to get greater correct and stable predictions. It constructs some of choice timber at training time and outputs the class this is the mode of the lessons output with the aid of person trees. It also tries to decrease the problems of excessive variance and high bias by means of averaging to discover a natural stability among the two extremes. both R and Python have strong packages to put into effect this algorithm.

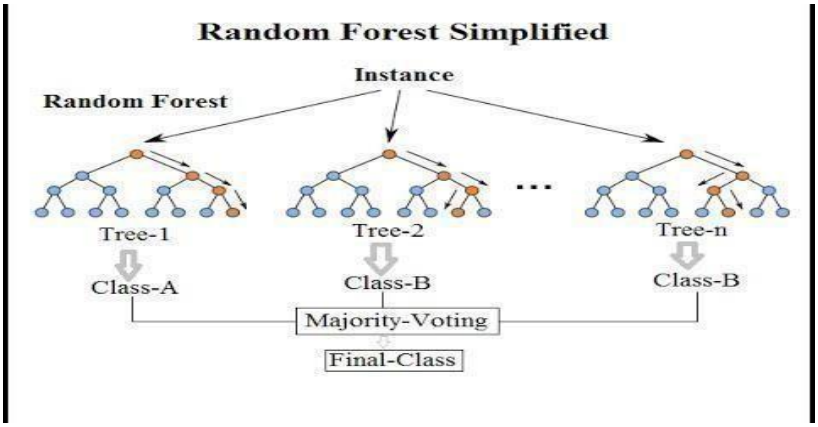


Figure:2.7.1.1 Random Forest Classifier

3.KNN:

KNN algorithm is one of the handiest classification algorithms and it's miles one of the most used getting to know algorithms. KNN is a non-parametric, lazy learning set of rules. Its reason is to use a dataset where in the facts factors are separated into several classes to expect the classification of a brand new sample point. A KNN algorithm uses a statistics and classifies new statistics points primarily based on a similarity measures (e.g. distance characteristic, mistakes price). type is done with the aid of a majority vote to its neighbours. The facts is assigned to the elegance which has the maximum nearest neighbours. As we boom the variety of nearest neighbours, the price of ok, accuracy can also increase.

When while we say a way is non-parametric, it way that it does now not make any assumptions on the underlying statistics distribution. In different phrases, the version structure is determined from the information. in case you consider it, it's quite useful, due to the fact in the "actual global", most of the information does no longer obey the everyday theoretical assumptions made (as in linear regression fashions, as an instance). therefore, KNN should and in all likelihood have to be one of the first selections for a class have a look at while there's little or no earlier information approximately the distribution data.

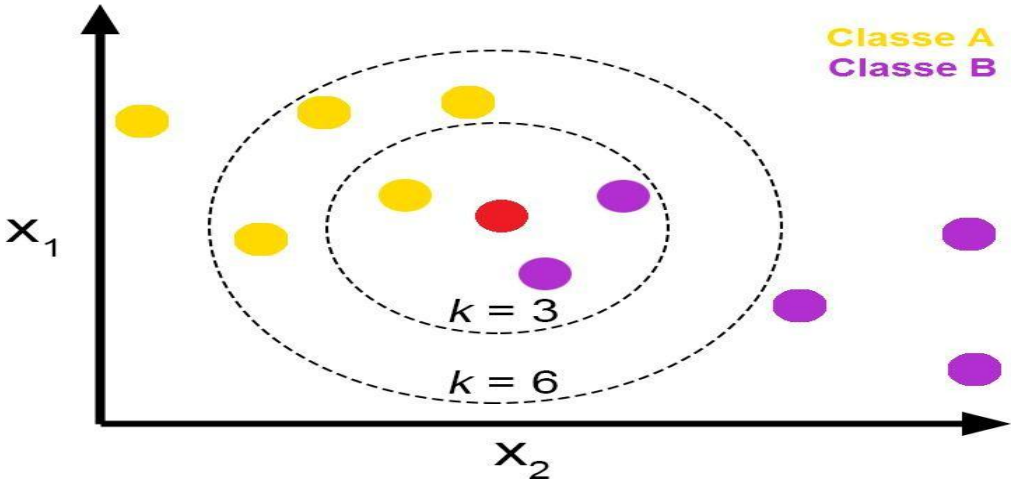


Figure:2.7.1.2 K-Nearest Neighbours

Logistic Regression:

Logistic regression is one of the maximum famous device studying algorithms, which comes beneath the Supervised studying approach. it is used for predicting the specific based variable the usage of a given set of independent variables.Logistic regression predicts the output of a express dependent variable. Consequently the outcome ought to be a specific or discrete cost. it is able to be both sure or No, zero or 1, actual or false, etc. but as opposed to giving the exact value as zero and 1, it offers the probabilistic values which lie among zero and 1.Logistic Regression is lots similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression issues, whereas Logistic regression is used for fixing the classification troubles

SMOTE Technique:

SMOTE (artificial minority oversampling method) is one of the most generally used oversampling methods to remedy the imbalance trouble. It ambitions to balance class

distribution by way of randomly increasing minority class examples by way of replicating them.SMOTE synthesis minority instances between existing minority instances. It generates the digital education data via linear interpolation for the minority class. those artificial schooling statistics are generated by way of randomly selecting one or more of the okay-nearest buddies for each instance inside the minority magnificence. After the oversampling technique, the information is reconstructed and several type models may be implemented for the processed information.

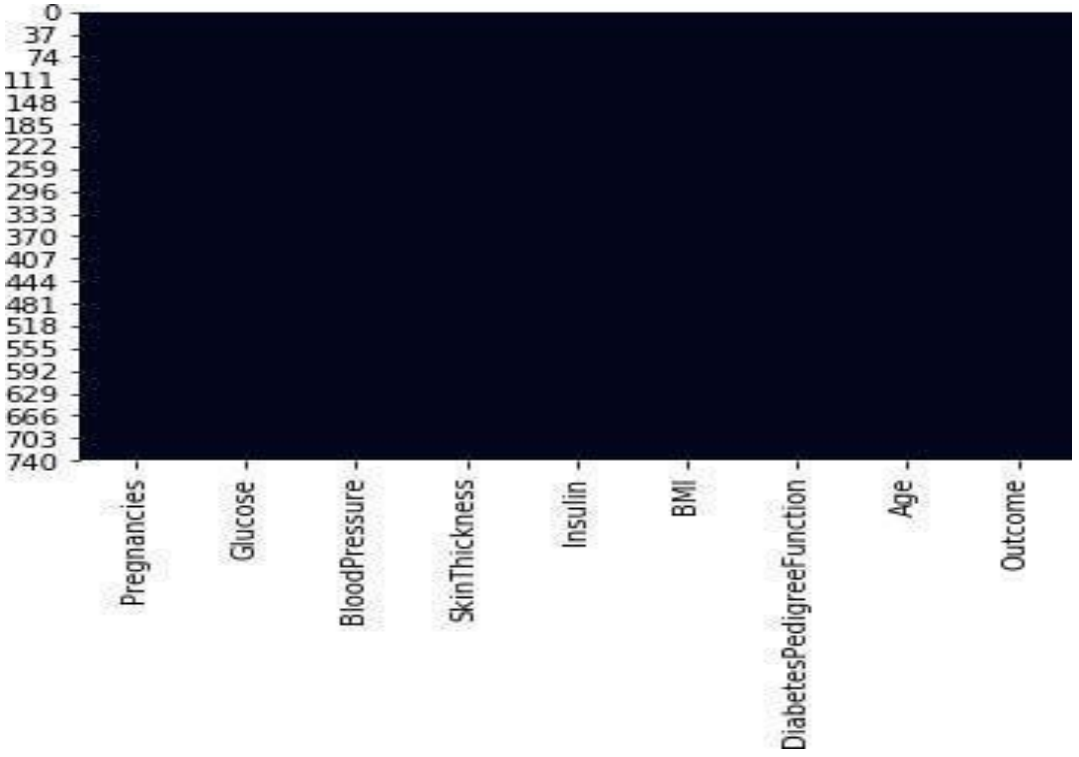
4.3.1 Replacing Missing Values

1. The statistics can have many irrelevant and missing parts. to deal with this element, statistics cleansing is carried out. It entails handling of missing facts, noisy statistics and so on.

2. This case arises whilst a few facts is lacking. it can be handled with the aid of filling the missing values manually, by way of attribute imply or the most in all likelihood cost. Filling missing values is one of the pre-processing techniques.

3. The lacking values in the data set is represented as '?' but it a non-standard missing cost and it must be converted into a fashionable lacking price Na N.

4. In order that pandas can locate the missing values. The fig1 under is a heatmap representing that there's no lacking values in the data set.

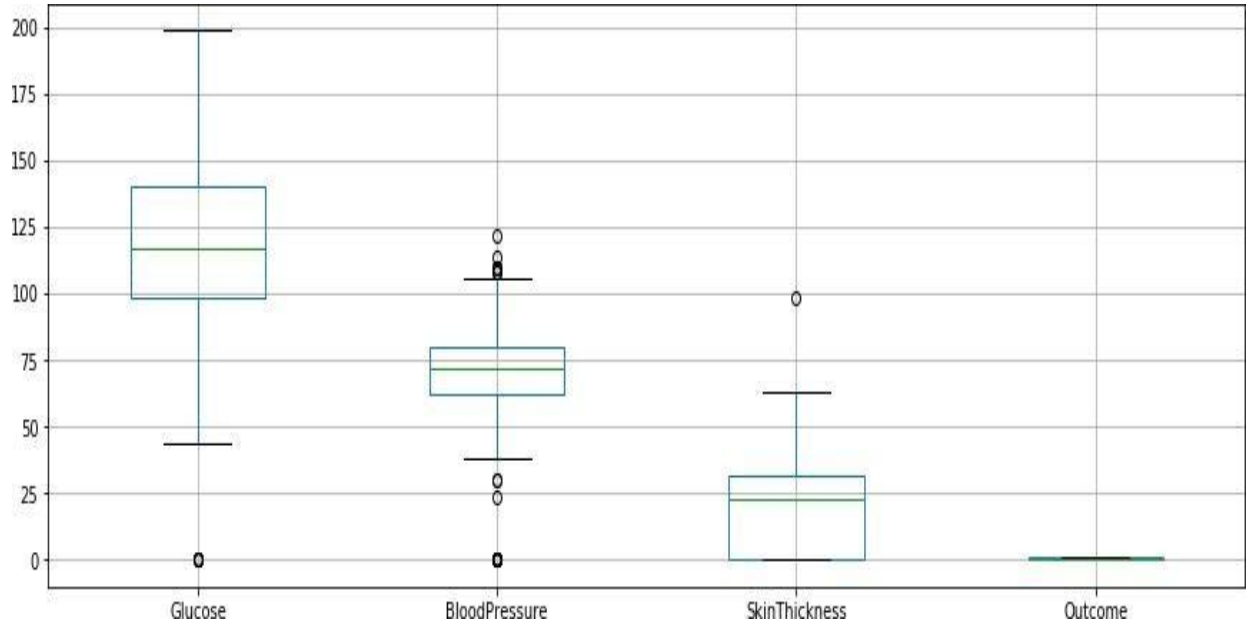


Here the formed graph predicts that there are no missing values in the given dataset.

4.3.2 Outliers Removal

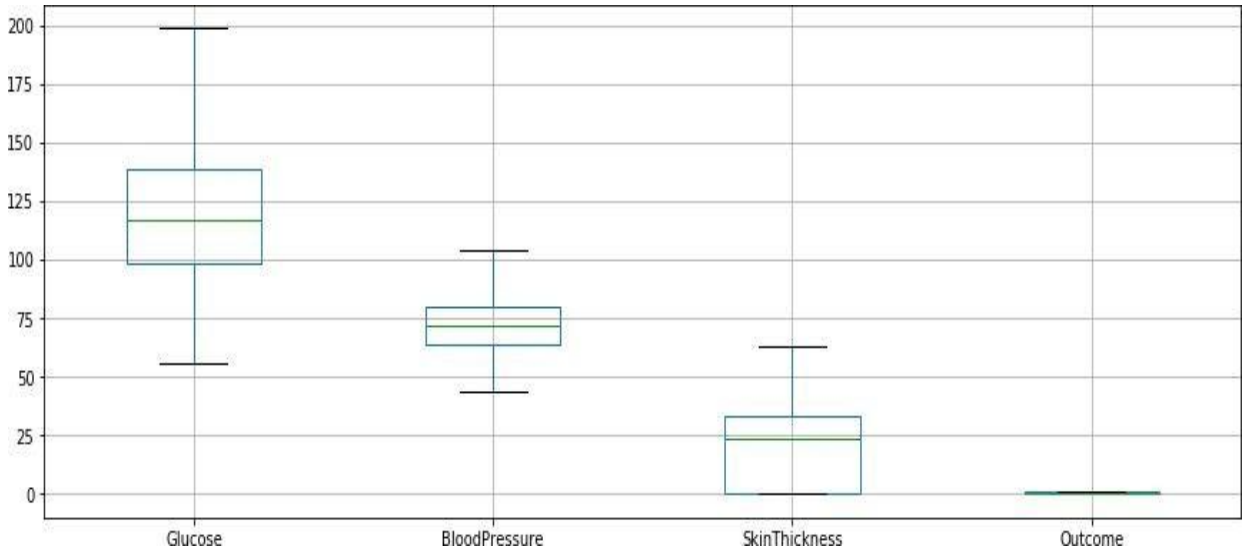
1. An outlier is an object that deviates extensively from the rest of the objects. They can be because of measurement or execution blunders.
2. The evaluation of outlier records is called outlier evaluation or outlier mining.
3. most facts mining techniques discard outlier noise or exceptions, however, in some applications such as fraud detection, the rare occasions may be extra exciting than the greater regularly happening one and therefore, the outier evaluation will become crucial in such case.

Before Outliers Removal



Here, the graph represents that there are some outliers in the dataset

After outlier removal



Here, the graph represents that there are no outliers in the dataset after outliers removal

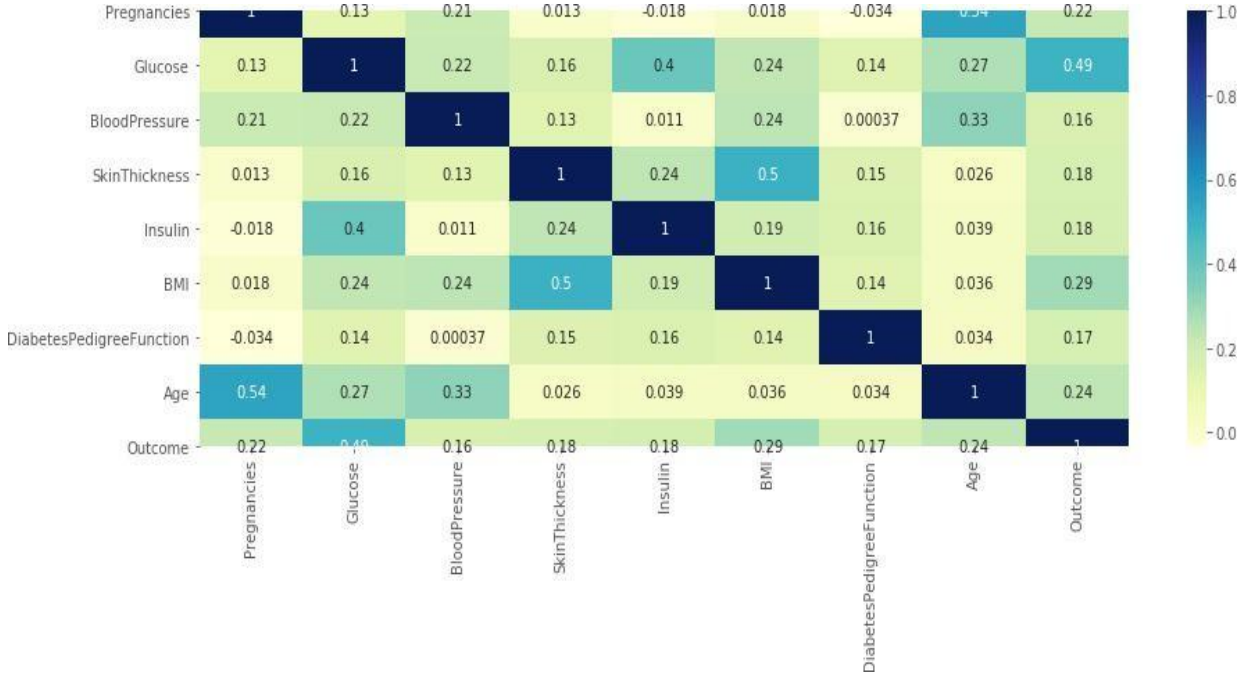
4.4 CORRELATION

1 .Correlation is a statistical degree that shows the extent to which two or extra variables vary collectively.

2.A fantastic correlation shows the quantity to which the ones variables boom or lower in parallel.

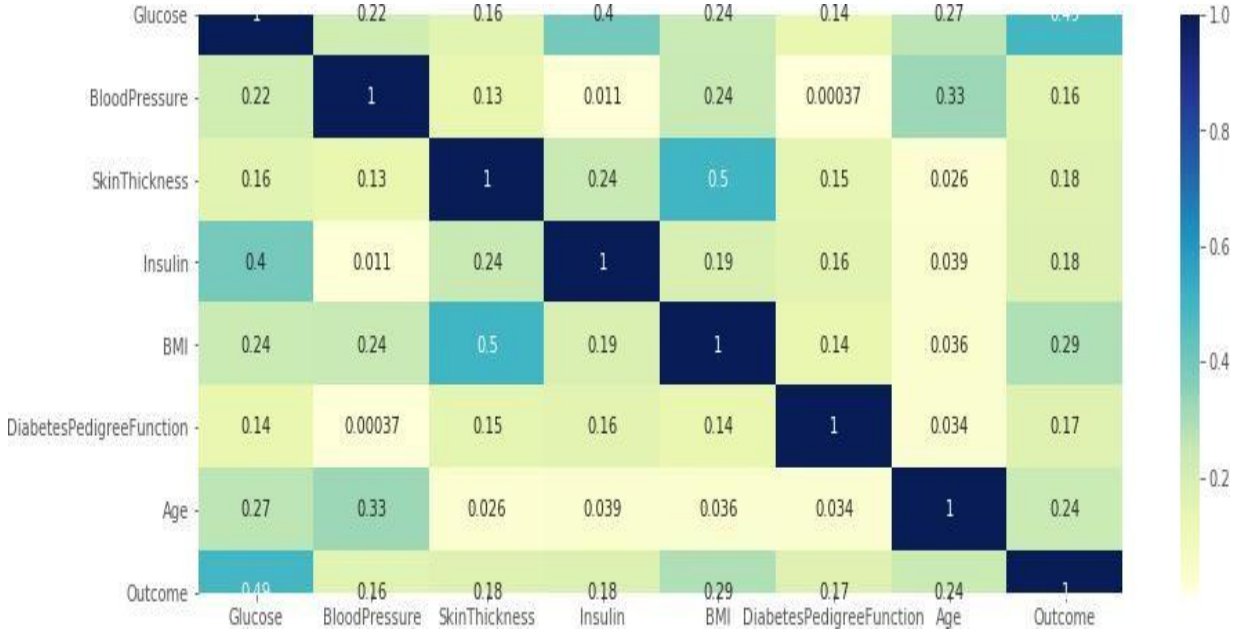
3..A poor correlation shows the volume to which one variable will increase as the alternative decreases.

Heatmap generated before correlation



Here, 0.54 is the maximum correlation value formed between age and pregnancies column.

Heatmap generated after correlation



Here, the figure represents the heatmap after dropping the pregnancies column.

4.5 CONFUSION MATRIX

A confusion matrix is a desk that is frequently used to explain the overall performance of a type version (or "classifier") on a hard and fast of test statistics for which the authentic values are known.

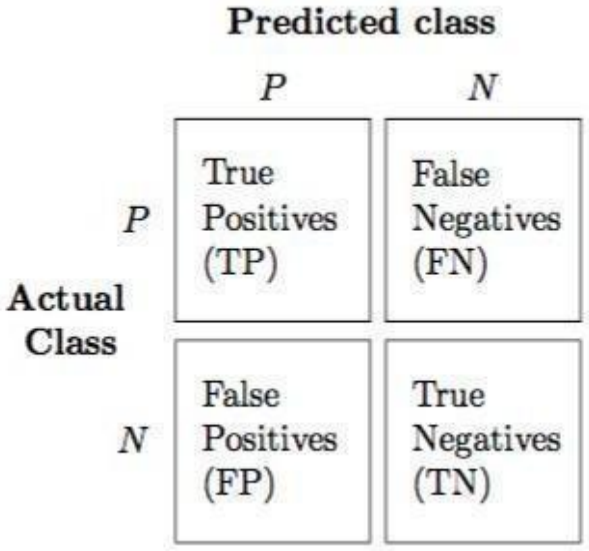


Fig:4.5.1 Confusion Matrix

A real high-quality (tp) is a result wherein the model predicts the superb class efficaciously. similarly, a real poor (tn) is an outcome in which the version efficaciously predicts the terrible class.

A false tremendous (fp) is an final results wherein the version incorrectly predicts the fine elegance. And a false terrible (fn) is an outcome where the version incorrectly predicts the negative magnificence.

Sensitivity or Recall or hit rate or true positive rate (TPR)

it's far the percentage of folks that definitely have the sickness were identified as having the sickness.

TPR = tp / (tp + fn)

Specificity, selectivity or true negative rate (TNR)

it's far the percentage of folks who sincerely do not have the sickness had been identified as not having the ailment.

TNR = tn / (tn + fp) =1-FPR

Precision or positive predictive value (PPV)

If the test end result is advantageous what's the chance that the patient truly has

the ailment.

PPV = tp / (tp + fp)

Negative predictive value (NPV)

If the take a look at result is terrible what's the chance that

the patient does not have disease

. NPV = tn / (tn + fn)

Miss rate or false negative rate (FNR)

If the check end result is horrible what is the risk that the patient does now not have disease

FNR = fn / (fp + tn)

Fall-out or false positive rate (FPR)

it's far the proportion of all the those who do now not have the sickness who will be identified as having the sickness

FPR = fp/ (fp + tn)

False discovery rate (FDR)

it's far the proportion of all the human beings diagnosed as having the disorder who do no longer have the ailment.FDR = fp / fp + tp

False omission rate (FOR)

it's far the proportion of the individuals with a bad take a look at result for which the real condition is positive.

FOR = fn / (fn + tn)

Accuracy

The accuracy reflects the total proportion of people which can

be efficaciously categorized

. ACC = (tp + tn) / (tp + tn + fp + fn)

F1 score

it's far the harmonic suggest of precision and

sensitivity

F1 = 2tp / (2tp+ fp + fn)

RMSE Score

Root suggest square of the error that has happened among the test values and the predicted values.

RMSE = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(|\quad - \quad|)^2}$

5. IMPLEMENTATION CODE

Part-1 Importing Data

```
import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

from sklearn.model_selection import train_test_split

import seaborn as sns

from sklearn.linear_model import LogisticRegression

from sklearn.preprocessing import StandardScaler

from sklearn.tree import DecisionTreeClassifier

from sklearn.metrics import accuracy_score

from sklearn.svm import SVC

from sklearn.neighbors import KNeighborsClassifier

from sklearn.ensemble import RandomForestClassifier

import warnings
```

Part- 2 Importing Data

```
df=pd.read_csv("diabetes.csv")
```

Part- 3 Checking Null Values

```
df.isnull().sum()
```

Part-4 Pre-processing Techniques

```
X.replace(to_replace=0,value=X.mean(),inplace=True)

X

X.boxplot()

X.boxplot(figsize=(13,5))

plt.show()

from sklearn.model_selection import train_test_split

X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.25,random_state=101)

X_train

X_test

Y_train.value_counts()

# Standard scalar

from sklearn.preprocessing import StandardScaler

std=StandardScaler()

X_train_std=std.fit_transform(X_train)

X_test_std=std.transform(X_test)

X_train_std

X_test_std
```

Part-5 Applying Algorithms

```
# Logistic Regression

from sklearn.linear_model import LogisticRegression

lr=LogisticRegression()
```

```
lr

lr.fit(X_train_std,Y_train)

Y_predict=lr.predict(X_test_std)

Y_predict

X_test

Y_test

from sklearn.metrics import accuracy_score

import pickle

log_score=accuracy_score(Y_test,Y_predict)*100

log_score

with open('diabetes.pkl','wb') as f:

    pickle.dump(lr,f)

lr_model=pickle.load(open('diabetes.pkl','rb'))


# Decision Tree Classifier

from sklearn.tree import DecisionTreeClassifier

dt=DecisionTreeClassifier()

dt.fit(X_train_std,Y_train)

Y_predict=dt.predict(X_test_std)

Y_predict

Y_test

decision_score=accuracy_score(Y_test,Y_predict)*100
```

```
decision_score

# MLPC CLassifier

from sklearn.neural_network import MLPClassifier

mlp=MLPClassifier(hidden_layer_sizes=(8,8))

mlp.fit(X_train_std,Y_train)

Y_predict=mlp.predict(X_test_std)

Y_predict

mlpc_score=accuracy_score(Y_test,Y_predict)*100

mlpc_score

# import svc

from sklearn.svm import SVC

model=SVC(kernel='rbf')

model.fit(X_train_std,Y_train)

Y_predict=model.predict(X_test_std)

svc_score=accuracy_score(Y_test,Y_predict)*100

svc_score

# Random Forest Classifier

from sklearn.ensemble import RandomForestClassifier

classifier=RandomForestClassifier()

classifier.fit(X_train,Y_train)

Y_predict=classifier.predict(X_test_std)

rand_score=accuracy_score(Y_test,Y_predict)*100
```

```
rand_score

# KNN Model

from sklearn.neighbors import KNeighborsClassifier

clf = KNeighborsClassifier(n_neighbors=3)

clf.fit(X_train,Y_train)

print(clf.score(X_test,Y_test)*100)

Y_pred=clf.predict(X_test_std)

KNN_score=accuracy_score(Y_test,Y_pred)*100

KNN_score

# Linear Regression

from sklearn.linear_model import LinearRegression

lr=LinearRegression()

lr.fit(X_train,Y_train)

y_predict=lr.predict(X_test_std)

y_predict

lr_score=accuracy_score(Y_test,Y_predict)*100

lr_score

models = pd.DataFrame({'Model': ["Logistic Regression", "DecisionTreeClassifier", "SVM-Linear",

"RandomForestClassifier", "KNN","Linear Regression","MLPC Classifier" ],'Accuracy Score':

[log_score,decision_score,svc_score,rand_score,KNN_score,lr_score,mlpc_score]})

models.sort_values(by = 'Accuracy Score', ascending = False, ignore_index=True)

models=['LR','SVC','MLPC','RF','KNN"DTC','LIR']

acc=[]
```

```
acc.append(round(log_score,2))

acc.append(round(decision_score,2))

acc.append(round(svc_score,2))

acc.append(round(rand_score,2))

acc.append(round(KNN_score,2))

acc.append(round(lr_score,2))

acc.append(round(mlpc_score,2))

fig = plt.figure()

import matplotlib

matplotlib.style.use('ggplot')

ax = fig.add_axes([0,0,1,1])

plt.title("Accuracy Score of different models ")

ax.bar(models,acc,width=0.8,color=['yellow','pink','skyblue','brown','green','orange'])
```

Part-5 Flask Code

```
from flask import Flask,render_template,request
import pickle
import numpy as np
from sklearn.linear_model import Logistic
Regression app = Flask( name )
@app.route('/')
def home():
    return render_template('index.')
@app.route('/predict', methods =['POST'])
def predict():
    if request.method == 'POST':
        preg = request.form['pregnancies']
        glucose = request.form['glucose']
        bp = request.form['bloodpressure']
        skin = request.form['skinthickness']
        bmi = request.form['bmi']
        insulin = request.form['insulin']
        dpf = request.form['dpf']
        age = request.form['age']
        lr_model=pickle.load(open('diabetes.pkl','rb'))
        data = np.array([[preg,glucose,bp,skin,bmi,insulin,dpf,age]])
        my_prediction = lr_model.predict(data)[0]
        print(data)
        print(my_prediction)
```

```
    return render_template('result.', prediction= my_prediction) if

name_____== (' main '):

    app.run(debug = True)
```

6. RESULT ANALYSIS

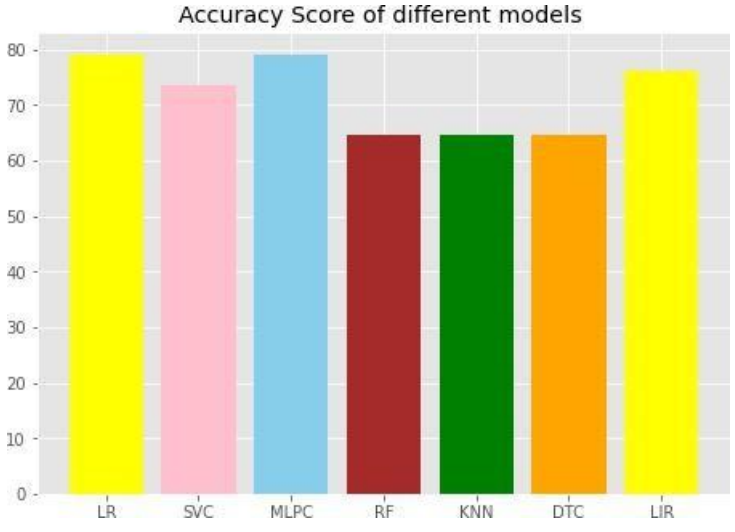


Figure:6.1 Comparison of accuracy of algorithms

Figure shows the comparison of accuracy ofsix classifiers (Logistic Regression, Support Vector Machine, Decision Tree Classifier, Random Factor Classifier, K-nearest Neighbors Classifier, Gaussian Naive Bayes). Support vector machine has 79.91% accurary which is quite helpful.

Figure shows the comparison of accuracy ofsix classifiers (Logistic Regression, Support Vector Machine, Decision Tree Classifier, Random Factor Classifier, K-nearest Neighbors Classifier, Gaussian Naive Bayes). There is no dropping of columns after using of chi-square method. 79.91% is the maximum accurary for logistic regression and vector machine.

7. TEST CASES

```
df=pd.read_csv("/diabetes (1).csv",engine='python')
```

```
df
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

768 rows × 9 columns

```
[7] df.head()
```

Fig.7.1 Output based on classifications

Fig.7.2 Output screen for Diabetes Disease Prediction

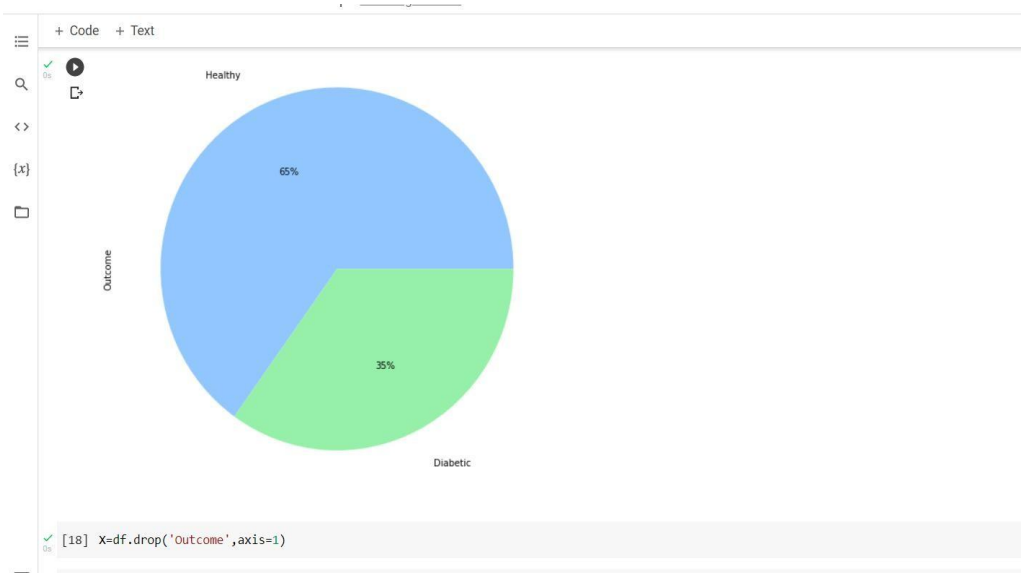
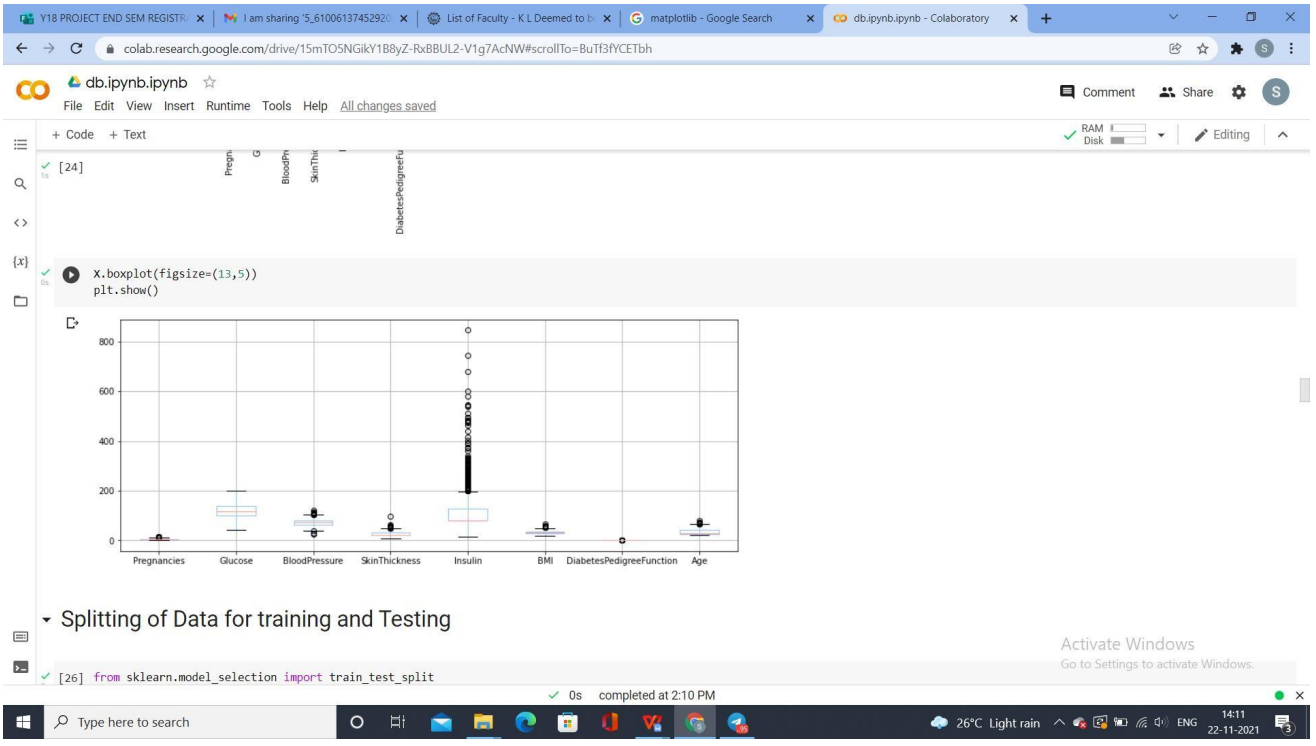


Fig.7.4 Splitting and testing



8. CONCLUSION

In this project we have used 6 algorithms for the proper implementation of the project to have the best accuracy of results.These algorithms gives the best output. The accuracy after applying feature selection are - The accuracy for Logistic Regression algorithm is 79.66%, The accuracy for Support Vector Classifier algorithm is 60.91%, The accuracy for Decision Tree algorithm is 69.09% , The accuracy for Random Factor algorithm is 79.45% ,The accuracy for KNN algorithm is 78.70% which is the highest accuracy.

9. FUTURE SCOPE

To develop our project in the form of android app or IOS app to provide the users to access the application in an easy manner.

10. REFERENCES

[1] https://www.kaggle.com/uciml/pima-indians-diabetes-database/data

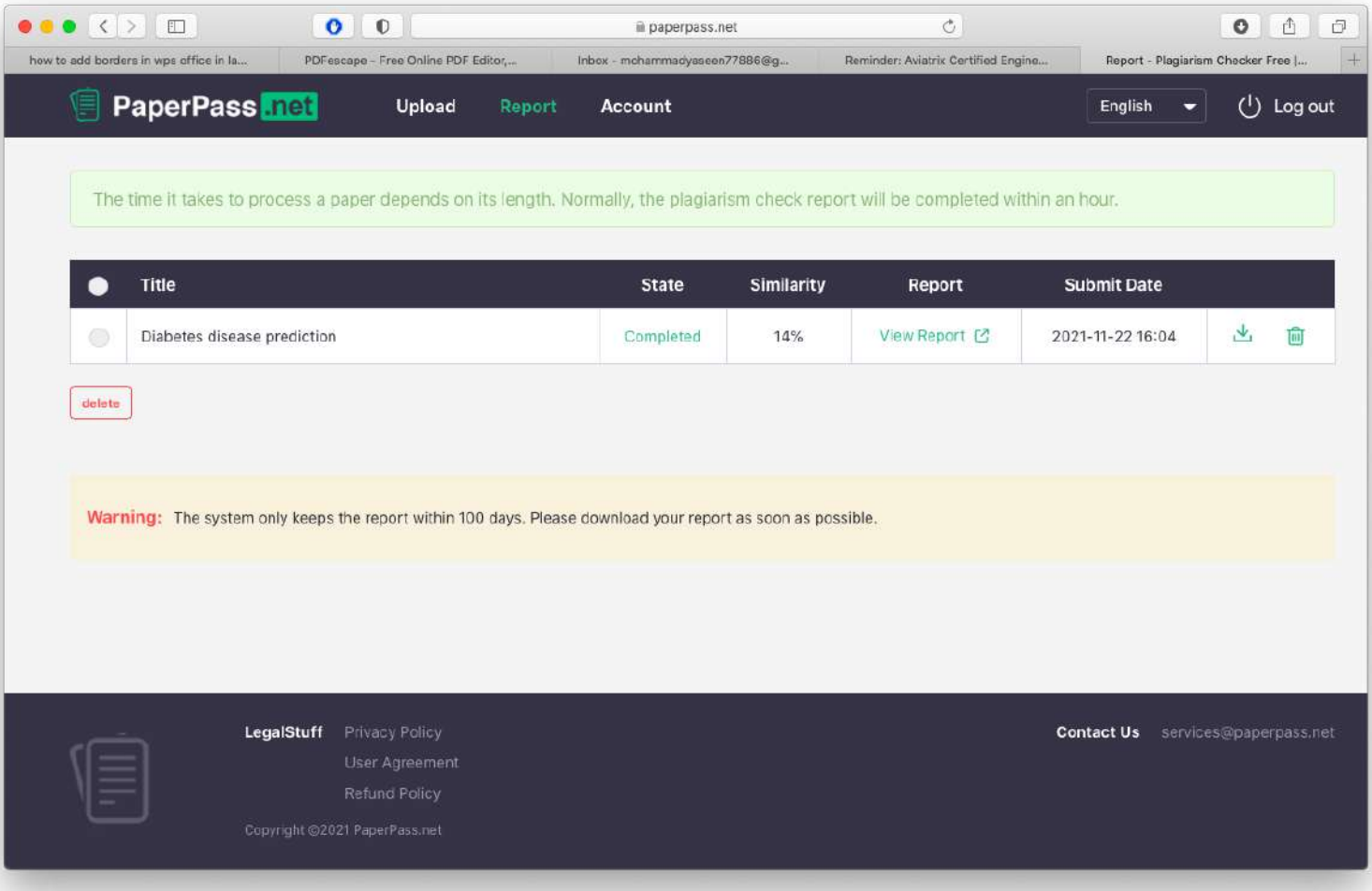
[2] https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine- learning-206b1f0b8223

[3] https://towardsdatascience.com/train-test-split-and-cross-validation-in-python- 80b61beca4b6

[4] https://machinelearningmastery.com/feature-selection-machine-learning-python/

[5] https://medium.com/@aneesha/recursive-feature-elimination-with-scikit-learn- 3a2cbdf23fb7

[6] https://medium.com/@saeedAR/smote-and-near-miss-in-python-machine-learning- in-imbalanced-datasets-b7976d9a7a79


paperpass.net
how to add borders in wps office in la...
PDFescape - Free Online PDF Editor,...
Inbox - mohammadyaseen77886@g...
Reminder: Aviatrix Certified Engine...
Report - Plagiarism Checker Free |...
PaperPass.net
Upload
Report
Account
English
Log out
The time it takes to process a paper depends on its length. Normally, the plagiarism check report will be completed within an hour.
Title
State
Similarity
Report
Submit Date
Diabetes disease prediction
Completed
14%
View Report
2021-11-22 16:04
delete
Warning: The system only keeps the report within 100 days. Please download your report as soon as possible.
LegalStuff
Privacy Policy
User Agreement
Refund Policy
Copyright ©2021 PaperPass.net
Contact Us
services@paperpass.net