Project Report

On

STATISTICAL STUDY ON DIABETES DATA

Submitted

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

APPLIED STATISTICS AND DATA ANALYTICS

by

NAFEESATH MISSRIYYA H M

(Register No. SM21AS013)

(2021-2023)

Under the Supervision of

RAJI S PILLAI

DEPARTMENT OF MATHEMATICS AND STATISTICS

ST. TERESA'S COLLEGE (AUTONOMOUS)

ERNAKULAM, KOCHI - 682011

MAY 2023

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

CERTIFICATE

This is to certify that the dissertation entitled, **STATISTICAL STUDY ON DIABETES DATA** is a bonafide record of the work done by Ms. **NAFEESATH MISSRIYYA H M** under my guidance as partial fulfillment of the award of the degree of **Master of Science in Applied Statistics and Data Analytics** at St. Teresa's College (Autonomous), Ernakulam affiliated to Mahatma Gandhi University, Kottayam. No part of this work has been submitted for any other degree elsewhere.

Date:
Place: Ernakulam

Raji S Pillai
Assistant Professor,
Department of Computer Application,
St. Teresa's College(Autonomous),
Ernakulam.

Ms Betty Joseph
Associate Professor & HOD,
Department of Mathematics and Statistics,
St. Teresa's College(Autonomous),
Ernakulam.

External Examiners

1:..... 2: .....

DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of **RAJI S PILLAI**, Assistant Professor, Department of Computer Applications, St. Teresa's College(Autonomous), Ernakulam and has not been included in any other project submitted previously for the award of any degree.

Ernakulam. **NAFEESATH MISSRIYYA H M**

Date: **SM21AS013**

ACKNOWLEDGEMENTS

I must mention several individuals who encouraged me to carry this work. Their continuous invaluable knowledgably guidance throughout the course of this study helped me to complete the work up to this stage

I am very grateful to my project guide Raji S Pillai for the immense help during the period of work

In addition, very energetic and competitive atmosphere of the Department had much to do with this work. I acknowledge with thanks to faculty, teaching and non-teaching staff of the department and Colleagues.

I also very thankful to HoD Ms.Betty Joseph for their valuable suggestions, critical examination of work during the progress.

Ernakulam. NAFEESATH MISSRIYYA H M

Date: **SM21AS013**

ABSTRACT

The best investment you ever make is in your health and one of the major factors that affects your health is the Lifestyle diseases. Among which the most common one is diabetes. Diabetes is one of the fastest growing life threatening disease that have already affected 422 million people worldwide as per the report of the World Health Organization 2018. After China, India is the country with the second-highest prevalence of diabetes, with an estimated 77 million people affected. India accounts for one in six people with diabetes worldwide. For a clinically significant outcome, early identification of diabetes is always preferred due to the existence of a relatively long asymptomatic phase. Around 50 percent of all people suffering from diabetes are undiagnosed because of its long term asymptomatic phase. So, it is important to study about the early symptoms of Diabetes and fit a model for the data. In this paper I am conducting exploratory data analysis on the diabetes data and comparing different machine learning models and finding the best model which can predict whether the patient is diabetic or not. Also finding the important factors that causes diabetes.

Contents

Chapter 1

Introduction

1.1 Diabetes

Diabetes is a rapidly growing disease among people, even among young people. To understand diabetes and how it develops, we need to understand what happens in the body without diabetes. Sugar (glucose) comes from the foods we eat, specifically carbohydrate foods. Carbohydrate foods provide our body with the main source of energy, everyone, even people with diabetes, needs carbohydrates. Carbohydrate foods include bread, cereals, pasta, rice, fruit, dairy products and vegetables.These foods are converted into glucose by the body when we eat them. The blood carries glucose around the body. Some of the glucose goes into our brain to help us think clearly and function. The rest of the glucose is taken to our body's cells for energy and also to our liver, where it is stored as energy for the body to use later. In order for the body to use glucose for energy, insulin is needed. The beta cells of the pancreas create the hormone insulin. Insulin acts as a key to a door. Insulin attaches to the door on the cell, opening the door to allow glucose to move from the bloodstream, through the door, and into the cell. If the pancreas is unable to produce enough insulin (insulin deficiency) or if the body is unable to use the insulin produced (insulin resistance), glucose builds up in the bloodstream (hyperglycaemia) and diabetes develops. Diabetes Mellitus means high levels of sugar (glucose) in the bloodstream and urine.

TYPES OF DIABETES

1. **TYPE 1 DIABETES : It is characterised by a weakened immune system and insufficient insulin production by cells. There are no convincing studies that demonstrate the causes of type 1 diabetes, and there are also no effective preventative measures at this time.**
2. **TYPE 2 DIABETES : It is characterised by either insufficient insulin production by the cells or improper insulin use by the body. 90 percent of people with diabetes have this kind of diabetes, making it the most prevalent type. Both genetic and lifestyle factors contribute to its occurrence.**
3. **GESTATIONAL DIABETES : Pregnant women who acquire high blood sugar quickly get gestational diabetes. After a pregnancy in which gestational diabetes was present, there is a high likelihood that type 1 or type 2 diabetes may develop.**

Diabetes Symptoms include frequent urination, increased thirst, fatigue, sleepiness, loss of weight, blurred vision, mood swings, confusion, and difficulty concentrating.

1.2 Objectives of the study

1. **To conduct the exploratory data analysis on the diabetes data.**
2. **To make predictions on whether a patient is diabetic or not based on the machine learning algorithms like Logistic regression, K Nearest Neighbour, Decision Tree and Random Forest Classifier and find the best model.**
3. **To find the early symptoms of diabetes**

Chapter 2

Literature Review

Results from related research that analysed various healthcare datasets and made predictions using a variety of methods and strategies are presented. Researchers have created and used a variety of prediction models utilising different data mining techniques, machine learning algorithms, or even a mix of these techniques.

- **Jyothi Rani, systematic attempts are undertaken to design a system that results in the prediction of diabetes in her paper, Diabetes prediction using machine learning. Five machine learning classification algorithms are examined and assessed in this paper using a variety of metrics. John Diabetes Database is the subject of experiments. Using the Decision Tree technique, experimental findings determine the suitability of the planned system with a 99 percent accuracy rate.**

- **Dr. Saravana Kumar N. M., Eswari, Sampath P., and Lavanya S. (2015) developed a system for the analysis of diabetic data using Hadoop and the Map Reduce approach. This approach forecasts the risk factors and type of diabetes. The technology is affordable for any healthcare business and is based on Hadoop.**

- **Aiswarya Iyer (2015) examined hidden patterns in a diabetes dataset using a classification technique. In this model, Naive Bayes and Decision Trees were employed. The usefulness of both algorithms was demonstrated as a consequence of a comparison of their performances.**

- **Humar Kahramanli and Novruz Allahverdi (2008) uses Artificial neural network (ANN) in combination with fuzzy logic to predict diabetes.**

- **B.M. Patil, R.C. Joshi and Durga Toshniwal (2010) introduced Hybrid Prediction Model which includes Basic K-means clustering approach, followed by application of classification algorithm to the result acquired from clustering algorithm. The C4.5 decision tree algorithm is used to create classifiers.**

Chapter 3

Methodology

3.1 Data Source

The data set in this paper comes from the open source standard test dataset website UCI (machine learning repository). The Sylhet Diabetes Hospital in Sylhet, Bangladesh, provided the data set, which was collected directly from 520 patients and approved by medical professionals.

3.2 Exploratory Data Analysis

Exploratory data analysis (EDA), a technique used in data mining, involves analyzing datasets to identify and summarize their key features, frequently with the help of data visualization techniques. Before beginning any modeling processes, EDA is used to determine what the data can reveal. Finding the characteristics in a table is not always simple; in these cases, an EDA can serve as a visual aid. EDA can be used to identify key characteristics and correlations between factors.

3.3 Factor Analysis

A process known as factor analysis is used to reduce a huge number of variables to a more manageable set of components. This method creates a common score by combining all variables' highest common variance. We can use this score as an index of all variables to do addi-

tional analysis. Factor analysis, a statistical technique used to describe variability among observed, correlated variables in terms of a possibly smaller number of unseen variables termed factors, is a component of general linear models (GLM). In response to unobserved latent variables, factor analysis looks for such joint fluctuations. Models for the observed variables include a "error" term and linear combinations of the potential factors. Finding independent latent variables is the goal of factor analysis. The assumption behind the factor analysis method is that it may be used to later minimise the number of variables in a dataset by learning more about the connections between observed variables.

3.3.1 Criteria for determining the number of factors

According to Kaiser Criterian, Eigen value is a good measure for determining a factor. If eigen value is greater than 1, we should consider that factor otherwise, that factor should not be considered.

3.3.2 Scree plot test

This method is to decide about the number of factors that should retain from the extracted factors. The test along with the plot determines which of the factors are actually contributing variance. The number of factors is plotted against proportion of variance. It extracts in the order of the extracted factors.

3.3.3 Kaiser Meyer Olkin Test(KMO Test)

It is the measure of sampling adequacy used to compare the magnitudes of the observed correlation coefficients in relation to the magnitude of the partial correlation coefficients. Large KMO values are good since correlation between pairs of variables can be explained by other variables. KMO value between 0.8 and 1 indicate that the sampling is adequate. KMO value less than 0.5 indicate that the sampling is not adequate and that necessary actions should be taken. KMO value close to zero means that there are large partial correlations compared to the sum of correlations.

3.4 Model building

3.4.1 Logistic Regression

Logistic regression is one of the Machine Learning algorithms that is most frequently used in the Supervised Learning group. Using a predetermined collection of independent variables, it is used to predict the categorical dependent variable. In a categorical dependent variable, the outcome is predicted by logistic regression. The result must be a discrete or categorical number. When classifying observations using various kinds of data, logistic regression can be used to quickly identify the variables that will work best. The dependent variable must be categorical in character, and the independent variable shouldn't be multi-collinear, according to the assumptions of logistic regression.
It is a mathematical function with the shape of the letter "S" that can transfer any real value to a range between 0 and 1. The sigmoid function is also known as the logistic function.
$f(z) = \frac{1}{1+e^{-z}}$ **where** $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_k X_k$
Parameter β_0 **is the intercept, Parameters** $\beta_1, \beta_2, ..., \beta_k$ **are referred as partial regression coefficients,** $X_1, X_2, ...X_k$ **are independent variables and z is the dependent variable.**

Hence, if the value of z increases to a positive infinity, the expected value of y will be 1, and if it decreases to a negative infinity, it will be 0. Also, if the result of the sigmoid function is greater than 0.5, the label is classified as class 1 or positive class, and if it is less than 0.5, it is classified as class 0 or negative class.

3.4.2 K Nearest Neigbour

Another supervised machine learning method is KNN. KNN aids in the solution of both classification and regression issues. KNN is a lazy prediction method. KNN implies that related things are located close to one another. Similar data sets are frequently found close together. KNN aids in classifying new work using a similarity metric. The KNN algorithm records every record and categorizes them based on how alike

they are. The algorithm locates the new data point's nearest neighbors, to make a prediction. Here, K is always a positive integer and stands for the number of close neighbors. The value of the neighbor is selected from a set of class n problems. Value of a neighbor is selected from a list of classes. Closeness determined in terms of Eucledian distance.
In Eucledian distance, the distance between the points p and q is given by
$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$.

3.4.3 Decision Tree

Decision trees are a machine learning technique that can be used to solve classification and regression problems. It is a graphical representation of a recursive partitioning procedure that divides a dataset into smaller subsets while maximizing the information gain or minimizing the impurity at each node. The tree is built using a top-down approach, starting with the root node, which contains the entire dataset, and splitting it into two or more child nodes based on a feature that provides the most significant discrimination between the classes. The process continues recursively until a stopping criterion is met, such as reaching a pre-defined depth, impurity threshold or minimum number of samples per leaf. The resulting tree can be interpreted as a sequence of if-else conditions that lead to a predicted outcome or value. Decision trees have several advantages, including their interpretability, ease of use, and ability to handle both categorical and continuous variables. However, they are prone to overfitting and can be sensitive to small changes in the data or structure. Therefore, various techniques such as pruning, ensembling, and regularization have been developed to improve their performance and robustness.

3.4.4 Random Forest Model

It belongs to the class of ensemble learning techniques and is employed in classification and regression applications. It is a type of ensemble method that combines multiple decision trees to produce a more accurate and robust model.

In a random forest, a large number of decision trees are generated, each of which makes its own prediction. The final prediction is determined by combining the predictions of all the trees. This approach helps to reduce overfitting and increase the accuracy of the model, as it averages out the predictions of individual trees and reduces the chance of any one tree making an erroneous prediction.

Each decision tree in the random forest is based on a random subset of the features and data points. This helps to ensure that each tree in the forest is unique and reduces the correlation between the trees, making the model more robust.

Random forest models have a number of advantages, including their ability to handle both categorical and numerical data, and their ability to handle missing and noisy data. They are also relatively easy to implement and interpret, making them a popular choice in machine learning applications.

3.4.5 Model Evaluation

Model evaluation is a critical process in machine learning that involves assessing the performance of a predictive model. It helps determine how well a model can generalize to new, unseen data and whether it can be trusted to make accurate predictions in real-world scenarios.

Confusion Matrix- It gives us a matrix as output and describes the complete performance of the model.

Accuracy - It is the ratio of number of correct predictions to the total number of input samples.

$Accuracy = \frac{TP+FN}{N}$ **where N= total number of predictions**

		Predicted class: P	Predicted class: N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Specificity - It is the Ratio of true negatives to total negatives in the data. $Specificity = \frac{TN}{TN+FP}$

Sensitivity - It is the ratio of true positive to the total positives in the data. $Specificity = \frac{TP}{TP+FN}$

ROC-AUC Curve - An indicator of performance for classification issues at different threshold levels is the AUC-ROC curve. AUC stands for the level or measurement of separability, and ROC is a probability curve. It reveals how well the model can differentiate across classes.Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1

Chapter 4

Dataset and Exploratory Data Analysis

4.1 Definitions of attributes

The dataset contains 17 variables.The response variable is the "Class" variable which gives the information whether the patient is diabetic or not. There are "age" and "gender" variables that specifies the Age and Gender of the patient. Rest of the 14 variables are the various early symptoms of Diabetes.

1. **Age: age of the patient**
2. **Gender: Gender of the patient**
3. **Polyuria: Whether the patient experienced excessive urination or not.**
4. **Polydipsia: Whether the patient experienced excessive thirst/ excess drinking or not.**
5. **Sudden weight loss: Whether the patient has an episode of sudden weight loss or not.**
6. **Weakness: Whether the patient has an episode of weakness.**
7. **Polyphagia: Whether the patient had an episode of excessive/ extreme hunger or not.**

8. **Genital thrust: Whether the patient had a yeast infection or not.**
9. **Visual blurring: Whether the patient had an episode of blurred vision.**
10. **Itching: Whether the patient had an episode of itch.**
11. **Irritability: Whether the patient had an episode of irritability.**
12. **Delayed Healing: Whether the patient had noticed delayed healing when wounded.**
13. **Partial Paresis: Whether the patient had an episode of weakening of a muscle/ group of muscles or not.**
14. **Muscle Stiffness: Whether the patient had an episode of muscle stiffness.**
15. **Alopecia: Whether the patient experienced hair loss or not.**
16. **Obesity: Whether the patient can be considered obese or not using his body mass index.**
17. **Class: Presence of Diabetes.**

Table 1 Description of attribute

	Attributes	Values
1	Age	16-90
2	Sex	1.Male, 0.Female
3	Polyuria	1.Yes, 0.No.
4	Polydipsia	1.Yes, 0.No.
5	Sudden weight loss	1.Yes, 0.No.
6	Weakness	1.Yes, 0.No.
7	Polyphagia	1.Yes, 0.No.
8	Genital thrush	1.Yes, 0.No.
9	Visual blurring	1.Yes, 0.No.
10	Itching	1.Yes, 0.No.
11	Irritability	1.Yes, 0.No.
12	Delayed healing	1.Yes, 0.No.
13	Partial paresis	1.Yes, 0.No.
14	Muscle stiffness	1.Yes, 0.No.
15	Alopecia	1.Yes, 0.No.
16	Obesity	1.Yes, 0.No.
17	Class	1.Positive, 0.Negative.

4.2 Sample Dataset

	age	gender	polyuria	polydipsia	sudden_weight_loss	weakness	polyphagia	genital_thrush	visual_blurring
0	40	Male	0	1	0	1	0	0	0
1	58	Male	0	0	0	1	0	0	1
2	41	Male	1	0	0	1	1	0	0
3	45	Male	0	0	1	1	1	1	0
4	60	Male	1	1	1	1	1	0	1

Figure 4.1:

itching	irritability	delayed_healing	partial_paresis	muscle_stiffness	alopecia	obesity	class
1	0	1	0	1	1	1	1
0	0	0	1	0	1	0	1
1	0	1	0	1	1	0	1
1	0	1	0	0	0	0	1
1	1	1	1	1	1	1	1

Figure 4.2:

Figure 4.1 and 4.2 given above shows the sample dataset

4.3 Data Pre-Processing

The dataset is checked for the null values. There was no null values in the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 520 entries, 0 to 519
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   age                 520 non-null    int64
 1   gender              520 non-null    object
 2   polyuria            520 non-null    int64
 3   polydipsia          520 non-null    int64
 4   sudden_weight_loss  520 non-null    int64
 5   weakness            520 non-null    int64
 6   polyphagia          520 non-null    int64
 7   genital_thrush      520 non-null    int64
 8   visual_blurring     520 non-null    int64
 9   itching             520 non-null    int64
 10  irritability        520 non-null    int64
 11  delayed_healing     520 non-null    int64
 12  partial_paresis     520 non-null    int64
 13  muscle_stiffness    520 non-null    int64
 14  alopecia            520 non-null    int64
 15  obesity             520 non-null    int64
 16  class               520 non-null    int64
dtypes: int64(16), object(1)
memory usage: 69.2+ KB
```

Figure 4.3:

4.4 Exploratory Data Analysis

We do some visualizations to obtain some basic insights from the dataset to give the ideas regarding the Dataset.

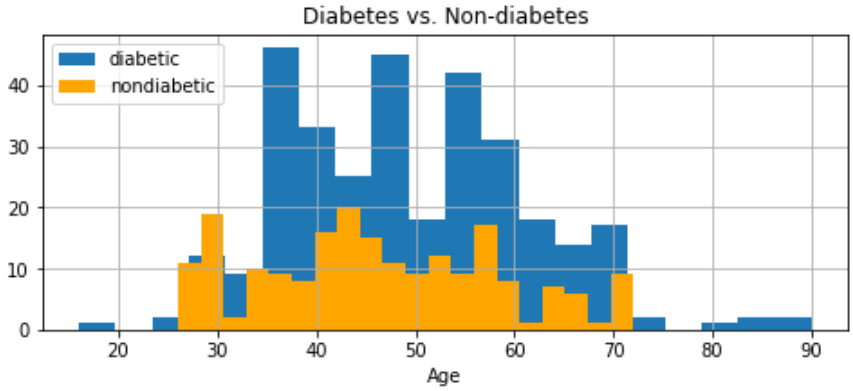

Figure 4.4:

in figure 4.4 we can see that people who have diabetes are heavily distributed between the age of 35 to 60 and are distributed over 80 years. However, those who don't have diabetes are distributed between 25 to 70 years old.

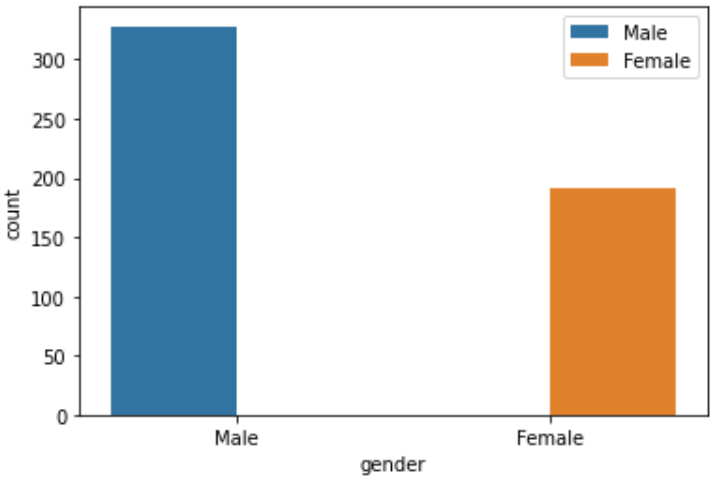

Figure 4.5:

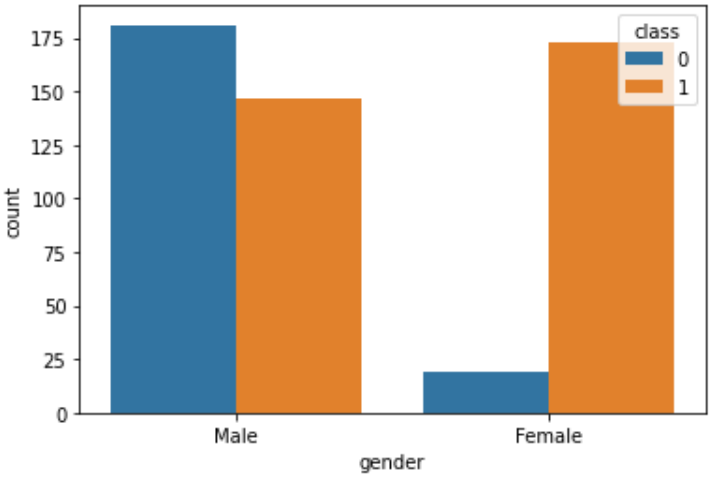

Figure 4.6:

Figure 4.5 shows the count plot. We can see that there is more male than females in the dataset.

Figure 4.6 shows the plot of diabetic and non diabetic males and females. Here we can see that more females are diagnosed with diabetes than males.

Chapter 5

RESULTS AND DISCUSSION

5.1 Factor Analysis

From the Total variance explained table and from the scree plot we can deduce the point that there are 4 factors for which the eigen values are greater than 1. So, they must be included in the model. So, a logistic regression model is built using these 4 factors as the independent variables and the class variable as the response variable. Its f1 score is also noted. The f1 score was 0.869, which is lesser as compared to that of full model. So, full model is considered to be better.

From figure 5.1,KMO value is between 0.8 and 1. So it is adequate

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.805
Bartlett's Test of Sphericity	Approx. Chi-Square	2571.688
	df	136
	Sig.	.000

Figure 5.1: KMO value

.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.359	25.642	25.642	4.359	25.642	25.642	3.603	21.194	21.194
2	2.372	13.953	39.595	2.372	13.953	39.595	2.239	13.168	34.363
3	1.511	8.888	48.483	1.511	8.888	48.483	2.011	11.828	46.191
4	1.208	7.106	55.589	1.208	7.106	55.589	1.598	9.398	55.589
5	.973	5.721	61.310						
6	.842	4.955	66.265						
7	.772	4.539	70.804						
8	.752	4.423	75.227						
9	.664	3.909	79.136						
10	.594	3.494	82.630						
11	.546	3.212	85.841						
12	.525	3.085	88.927						
13	.490	2.882	91.809						
14	.416	2.448	94.256						
15	.411	2.417	96.673						
16	.328	1.928	98.602						
17	.238	1.398	100.000						

Extraction Method: Principal Component Analysis.

Figure 5.2: Table explaining the total variance and eigen value of various factors

Component Matrix[a]

	Component			
	1	2	3	4
polydipsia	.743			
class	.734	-.376		
polyuria	.733			
partial_paresis	.668			
polyphagia	.604			
visual_blurring	.558		-.416	
sudden_weight_loss	.541		.309	
weakness	.528			
gender_auto	.470	-.390		
muscle_stiffness	.458	.337		.339
irritability	.377		.375	
alopecia		.745		
itching	.313	.598		-.319
age	.416	.559		
delayed_healing	.369	.549		-.449
genital_thrush			.774	
obesity				.679

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

Figure 5.3: fators divided into various components using PCA

Rotated Component Matrix[a]

	Component			
	1	2	3	4
class	.858			
polydipsia	.780			
polyuria	.777			
sudden_weight_loss	.667			
partial_paresis	.542	.370		-.308
gender_auto	.530			-.468
muscle_stiffness		.708		
visual_blurring		.698		
age		.613	.305	.312
polyphagia	.404	.443		
delayed_healing			.783	
itching			.711	
weakness	.327		.498	
alopecia	-.451		.455	.437
genital_thrush		-.344		.697
irritability	.307			.463
obesity		.422	-.384	.459

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 13 iterations.

Figure 5.4: faCtors divided into components after rotation

Component Transformation Matrix

Component	1	2	3	4
1	.814	.498	.299	.008
2	-.495	.401	.669	.383
3	.273	-.450	-.015	.850
4	-.135	.623	-.680	.361

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

Figure 5.5: component transformation matrix

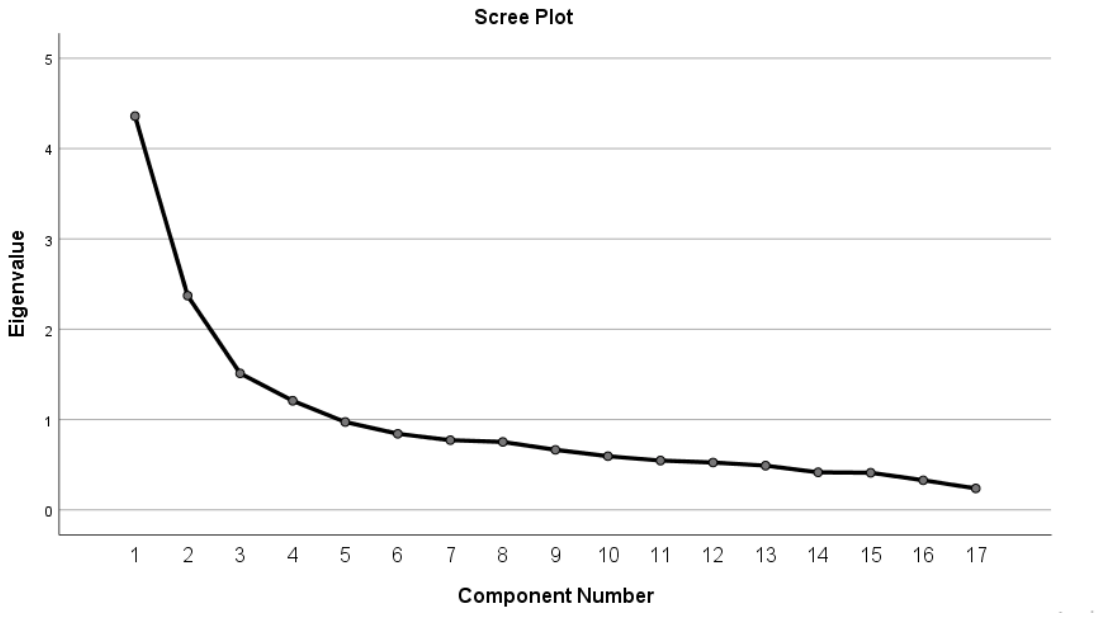

Figure 5.6: Scree plot

5.2 Logistic Regression Model

A logistic regression model was fitted firstly with those 4 factors that were found significant using factor analysis, then the full model. Accuracy are compared.

It was found that full model has more accuracy of 91 percent as compared to reduced model which only has an accuracy of 86 percent. Therefore, the full model is better.

Confusion matrix is given by,

75 8

11 114

Accuracy: 0.9086538

Sensitivity: 0.903614

Specificity: 0.912

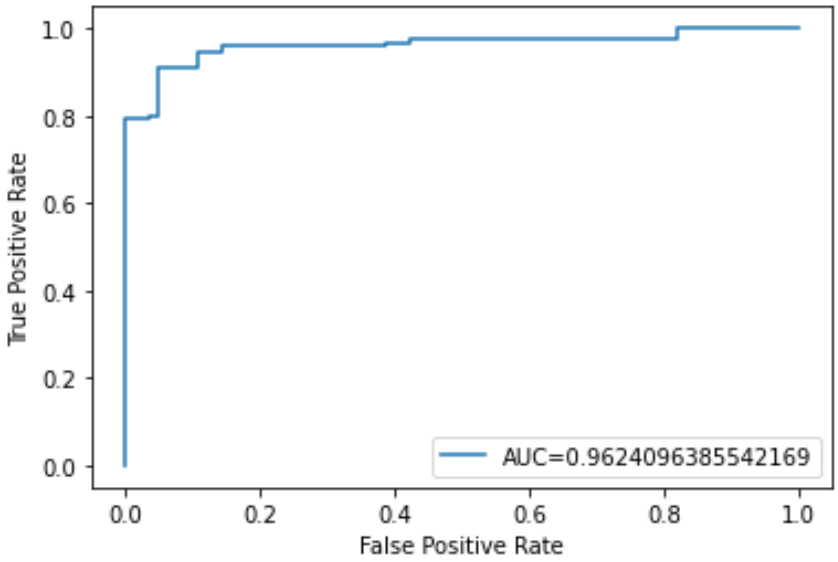

Figure 5.7: ROC-AUC for logistic regression

5.3 KNN

Confusion matrix is given by,

79 4

10 115

Accuracy: 0.9326923

Sensitivity: 0.951807

Specificity: 0.92

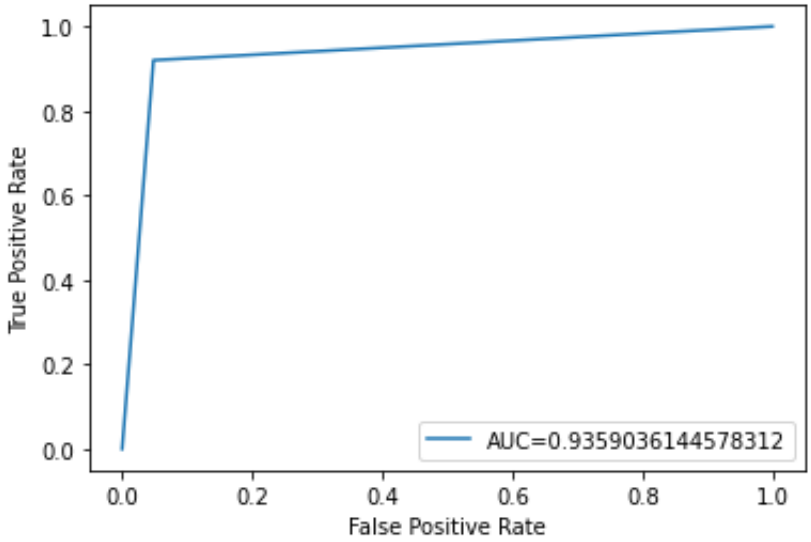

Figure 5.8: ROC-AUC for KNN

5.4 Decision Tree

Confusion matrix is given by,

75 8

5 120

Accuracy: 0.9375

Sensitivity: 0.903614

Specificity: 0.96

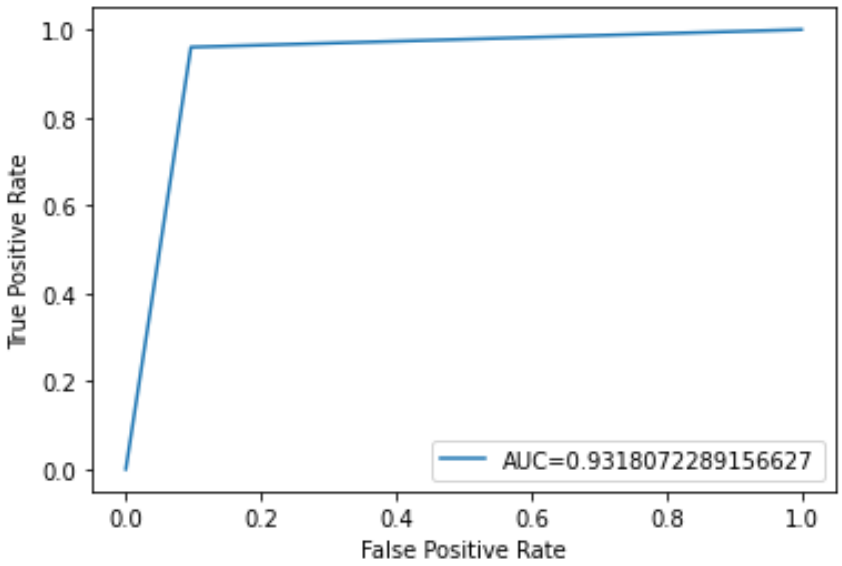

Figure 5.9: ROC-AUC for Decision Tree

5.5 Random Forest Classifier

Confusion matrix is given by,

81 2

3 112

Accuracy: 0.97596

Sensitivity: 0.9759

Specificity: 0.976

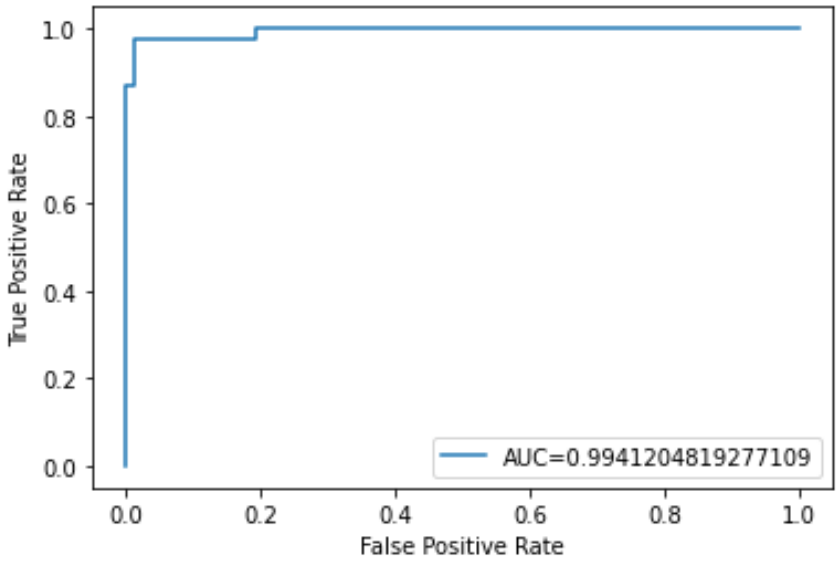

Figure 5.10: ROC-AUC for Random Forest Classifier

5.6 Comparison of Models

Different classification algorithms were applied to our dataset, and results of all techniques were slightly different as the working criteria of each algorithm is different. The accuracy of different models are compared and we can see from Figure 5.11 that random forest shows maximum accuracy of 97 percent as compared to other models.

To visualise the performance of the model we plot the ROC-AUC curve for different models. AUC value of different models are shown in Figure 5.11 and we can see that Random forest has the highest AUC value of 0.99 which is so close to 1. So we can say that Random Forest classifier is the best model for predicting whether the person is diabetic or not.

Model	Accuracy	AUC Value
Logistic regression	91%	0.96
K nearest neighbour	93%	0.93
Decision tree	94%	0.93
Random forest	97%	0.99

Figure 5.11: comparison of models

5.7 Feature Importance plot

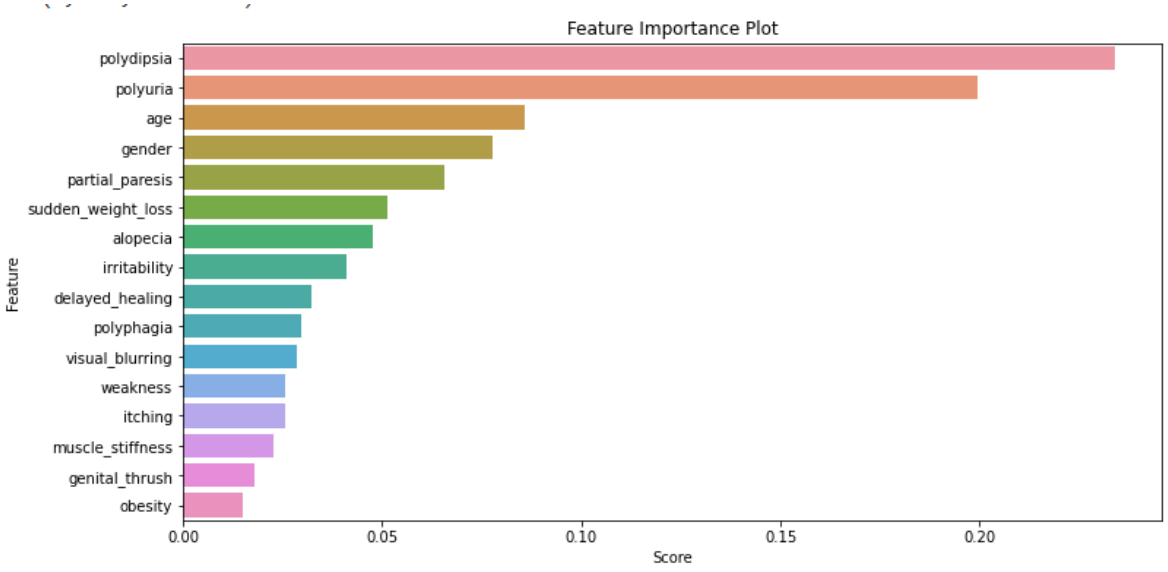

Figure 5.12:

We get random forest classifer as the best model for predicting whether a patient is diabetic or not. The important features taken for prediction of diabetes in random forest is shown in figure 5.12. We can see that polydipsia and polyuria as are 2 important conditions which help to predict a person having diabetes. Polydipsia is the medical term for experiencing extreme thirst while Polyuria is a condition where a person has excessive urination.

Chapter 6

Conclusion

One of the important real-world medical problem is the detection of diabetes at its early stage.In this, study, systematic efforts are made in designing a system which result in the prediction of diabetes. Four machine learning classification algorithms are studied and evaluated on various measures during the work. Experiments are performed on the early stage risk prediction dataset from UCI (machine learning repository). Experimental results determine the adequacy of the designed system with an achieved accuracy of 97 percent using the Random Forest algorithm.

In the future, the designed system with machine learning classification algorithm can be used to predict or diagnose diabetes. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.

Chapter 7

References

1. Talha Mahboob Alam, Muhammad Atif Iqbal, Yasir Ali, Abdul Wahab, Safdar Ijaz, Talha Imtiaz Baig, Ayaz Hussain, Muhammad Awais Malik, Muhammad Mehdi Raza, Salman Ibrar, Zunish Abbas, 2019, A model for early predition of diabetes, Informatics in Medicine Unlocked, Volume 16, 100204, ISSN 2352-9148..

2. Mitushi Soni , Dr. Sunita Varma, 2020, Diabetes Prediction using Machine Learning Techniques, international journal of engineering research technology (ijert) Volume 09, Issue 09.

3. Rani, KM. (2020). Diabetes Prediction Using Machine Learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology.
294-305. 10.32628/CSEIT206463.

4. Mujumdar, Aishwarya, and V Vaidehi. "Diabetes Prediction Using Machine Learning Algorithms." Procedia Computer Science, vol. 165, 2019, pp. 292–299.

5. Sahoo, Sipra; Mitra, Tushar; Mohanty, Arup Kumar; Sahoo, Bharat Jyoti Ranjan; and Rath, Smita (2022) "Diabetes Prediction: A Study of Various Classification based Data Mining Techniques," International Journal of Computer Science and Informatics: Vol. 4 : Iss. 3 , Article 1.

6. "UCI Machine Learning Repository: Early Stage Diabetes Risk Prediction Dataset. Data Set." Archive.ics.uci.edu,

7. Ghosh, Soumyajit. (2022) ,"Diabetes-Prediction." GitHub, github.com/SoumG/Diabetes-Prediction.

8. Khare, Sangita.(2020) "Diabetes Prediction Using Machine Learning Techniques." HELIX, www.academia.edu/88329279/Diabetes-Prediction-using-Machine-Learning-Techniques. Accessed 28 Jan. 2023.