# PREDICTION OF DIABETES USING MACHINE LEARNING

AV Srinivas
Assistant Professor
CSE (Data Science) Dept.
Sreyas Institute of Engineering and Technology, Telangana, India
av.srinivas@sreyas.ac.in

Abbireddy Ramya
CSE, Sreyas Institute of Engineering and Technology, Telangana, India
abbireddyramya1994@gmail.com

GT Chandralekha
CSE, Sreyas Institute of Engineering and Technology, Telangana, India
chandralekha.gt@gmail.com

Bhandaram Vaagdevi
CSE, Sreyas Institute of Engineering and Technology, Telangana, India
vaagdevibhandaram45@gmail.com

K Anand Goud
CSE, Sreyas Institute of Engineering and Technology, Telangana, India
anandgoud7678@gmail.com

***Abstract- Diabetes mellitus could be a gathering of metabolic maladies wherever aldohexose levels area unit overly high. About 8.8% of the planet was diabetic in 2017. it's anticipated that this can reach nearly 100 percent by 2045. the numerous checks are that once AI based mostly classifiers area unit applied to such informational indexes for likelihood stratification, prompts lower execution. during this manner, our goal is to create up a efficient and vigorous Machine Learning (ML) framework below the presumption that missing qualities or anomalies whenever supplanted by a middle style can yield higher hazard stratification exactitude.***

***Keywords-****Machine Learning, Diabetes mellitus, Aldohexose, Stratification*

## I. INTRODUCTION

Diabetes is one of the most hazardous diseases on the planet. Diabetes is caused by obesity, excessive blood glucose levels, and other factors. It alters the hormone insulin, causing aberrant crab metabolism and improving blood sugar levels. When the body does not produce enough insulin, diabetes develops. According to the World Health Organization, 422 million people worldwide suffer from diabetes, with the majority living in low- or middle-income nations. Up until 2030, this figure might be boosted to 490 billion. Diabetes is, nevertheless, prevalent in a number of countries, including Canada, China, and India. With a population of more than 100 million people, India has a total of 40 million diabetes. While we didn't achieve our goal of 100 percent accuracy in diabetes prediction, we did develop a system that can come close to it given enough time and data. As with any project of this nature, there is room for improvement. Because of the nature of this project, multiple algorithms can be combined as modules and their results combined to improve the accuracy of the final result. This research could be expanded to see how likely non-diabetic people are to develop diabetes in the

coming years. Thus, for this purpose we apply popular classification and ensemble methods on dataset for prediction. Diabetes is a common chronic disease that threat to human health. As a result, we use common classification and ensemble algorithms on the dataset to make predictions. Diabetes is a prevalent chronic disease that can be extremely dangerous to one's health. Diabetes is diagnosed when blood glucose levels are greater than normal, which is caused by insulin secretion or biological factors. Diabetes can harm our bodies in a variety of ways, including causing tissue, kidney, eye, and blood artery dysfunction. Based on physical examination data and consultation with doctors, machine learning may make a preliminary diagnosis of diabetes mellitus. Many techniques, including machine learning methods like Random Forest, Support Vector Machine, Decision Tree, and others, have recently been utilised to predict diabetes. We can forecast diabetes using machine learning approaches by creating predicting models based on medical datasets.

## II. LITERATURE SURVEY

[1] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference on I˜SMAC,978-1-5090-3243-3,2017**.**

Nowadays, the health-care industry generates a tremendous amount of data. To extract knowledge from this data and use it to make important decisions, it must be collected, stored, and processed. Diabetic Mellitus (DM) is a noncommunicable disease (NCD) that affects a large number of people. DM has now become a major health concern in emerging countries like India. Diabetes mellitus (DM) is a serious disease with long-term consequences and a slew of health issues. It is vital to develop a system that can save and evaluate diabetic data and identify potential problems using technology. Predictive analysis is a process that combines multiple data mining techniques, machine learning algorithms, and statistics to obtain insight and anticipate future risks using current and historical data sets. In this work, a machine learning technique is used in a Hadoop MapReduce context to find missing values in a Pima Indian diabetes data set and to discover trends. This work will be able to forecast common forms of diabetes, associated future hazards, and the sort of treatment that can be supplied based on the patient's risk level.

**[**2] Ayush Anand and Divya Shakti," Prediction of Diabetes Based on Personal Lifestyle Indicators", 1st International Conference on Next Generation Computing Technologies, 978-1- 4673-6809-4, September

Diabetes Mellitus is one of the most serious diseases, and it affects a large number of people. Diabetes Mellitus can be caused by age, obesity, lack of exercise, inherited diabetes, lifestyle, poor food, high blood pressure, and other factors. Diabetes increases the risk of ailments such as heart disease, renal disease, stroke, vision problems, nerve damage, and so on. The current hospital practise is to collect required information for diabetes diagnosis through various tests, and then give suitable therapy depending on the diagnosis. In the healthcare industry, big data analytics is extremely important. The healthcare industry has vast datasets. Big data analytics can be used to examine large datasets and uncover hidden information and trends in order to gain knowledge from the data and forecast outcomes. The categorization and prediction accuracy of the existing approach is not very good. In this study, we offer a diabetes prediction model for better diabetes classification that combines a few external factors that cause diabetes, as well as regular components such as glucose, BMI, age, and insulin. When compared to the old dataset, the new dataset improves classification accuracy. Furthermore, a diabetes prediction pipeline

model was imposed with the goal of boosting classification accuracy.

[3] B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 978-1-5386- 2745-7,2017.

Machine learning is the way to go when we have a large data collection on which we want to perform predictive analysis or pattern identification. Machine Learning (ML) is the fastest-growing field in computer science, and health informatics is a particularly difficult problem to solve. Machine Learning aims to create algorithms that can learn and improve over time and be used to make predictions. Machine learning techniques are widely used in a variety of fields, and machine learning prediction techniques have benefited the health care industry in particular. It provides a number of alerting and risk management decision-making tools aimed at enhancing patient safety and healthcare quality. With the desire to cut healthcare expenses and a shift toward customised care, the healthcare industry is changing. The healthcare business faces hurdles in critical areas such as electronic record management, data integration, and computer aided diagnostics and disease predictions as a result of the need to cut healthcare costs and the shift toward individualised treatment. To address these issues, machine learning provides a variety of tools, methodologies, and frameworks. This paper presents a study of several Machine Learning prediction methodologies and tools in practise. A look at Machine Learning's applications in many fields is also covered, with a focus on its importance in the health-care industry.

[4] Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing,2015

The complexity of modernising the healthcare industry's trend toward processing enormous health records and accessing them for analysis and action will considerably rise. Because of the unstructured nature of Big Data in the health business, it is required to structure and emphasise its magnitude into a nominal value with a feasible solution. The healthcare industry has numerous obstacles, which highlights the relevance of data analytics development. Diabetic Mellitus (DM) is a noncommunicable disease that is a substantial health risk in developing countries like India. The acute character of DM is linked to a variety of long-term consequences and health problems. In this study, we employ a Hadoop/Map Reduce environment and a predictive analysis method to forecast the most common diabetes kinds, their complications, and the type of treatment to be given. This approach, according to the report, provides an effective way to cure and care for patients with superior outcomes such as affordability and availability...

[5] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly," Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.

Diabetes affects around 246 million people globally, with women accounting for the majority of cases. According to the WHO, this figure is predicted to reach over 380 million by 2025. With no solution in sight, the disease has been dubbed the fifth deadliest disease in the United States. Diabetes cases and symptoms are widely recorded, thanks to the emergence of informative technology and its continuing entry into the medical and healthcare sectors. This research tries to find

solutions to diagnose the condition by using Decision Tree and Nave Bayes algorithms to analyse the patterns revealed in the data through classification analysis. The goal of the study is to develop a faster and more efficient method of identifying the disease, allowing patients to receive treatment sooner.

## III. EXISTING SYSTEM

Diabetic identification is still done manually by doctors nowadays. The accuracy of which cannot be thoroughly checked. According to surveys, anywhere from 30 to 80 percent of diabetic illnesses go undetected, and many incorrect diagnoses occur

**Disadvantages:**

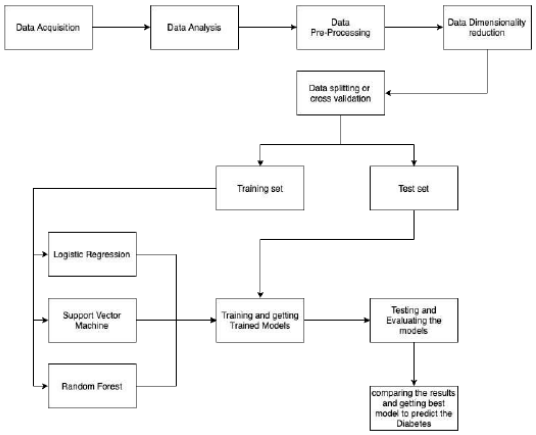
• The assessment is still done manually, resulting in many inaccuracies and incorrect treatments.

## IV. PROPOSED SYSTEM

Python-based Machine Learning Algorithm has been trained on Diabetic Patients Database in our proposed system. Patients and professionals can use the Visualized Decision Tree to analyse Diabetes with ease.

**Advantages**:

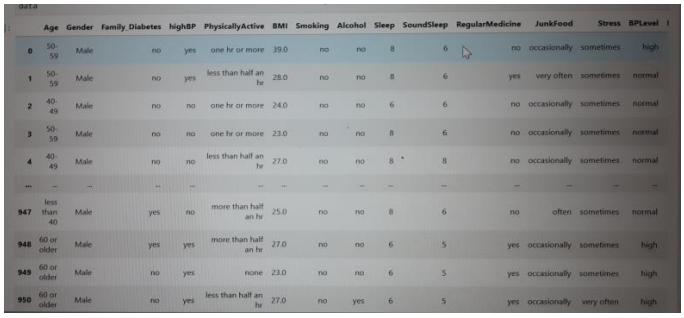
- Higher accuracy can be achieved with less training data.
- Use of a Machine Learning Algorithm improves the system's reliability and accuracy.



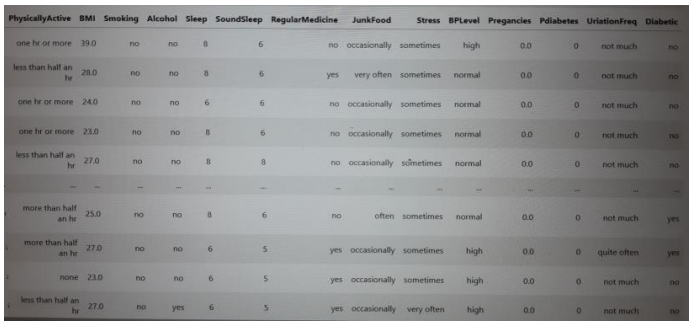

System Architecture

## V. DATA DESCRTPION AND RESULTS

https://www.kaggle.com/johndasilva/diabetes provided the diabetes data set. Diabetes database with 952 cases. The goal is to determine whether or not the patient is diabetic based on the measurements.



| | Age | Gender | Family_Diabetes | highBP | PhysicallyActive | BMI | Smoking | Alcohol | Sleep | SoundSleep | RegularMedicine | JunkFood | Stress | BPLevel |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 50-59 | Male | no | yes | one hr or more | 39.0 | no | no | 8 | 6 | no | occasionally | sometimes | high |
| 1 | 50-59 | Male | no | yes | less than half an hr | 28.0 | no | no | 8 | 6 | yes | very often | sometimes | normal |
| 2 | 40-49 | Male | no | no | one hr or more | 24.0 | no | no | 6 | 6 | no | occasionally | sometimes | normal |
| 3 | 50-59 | Male | no | no | one hr or more | 23.0 | no | no | 8 | 6 | no | occasionally | sometimes | normal |
| 4 | 40-49 | Male | no | no | less than half an hr | 27.0 | no | no | 8 | 8 | no | occasionally | sometimes | normal |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 947 | less than 40 | Male | yes | no | more than half an hr | 25.0 | no | no | 8 | 6 | no | often | sometimes | normal |
| 948 | 60 or older | Male | yes | yes | more than half an hr | 27.0 | no | no | 6 | 5 | yes | occasionally | sometimes | high |
| 949 | 60 or older | Male | no | yes | none | 23.0 | no | no | 6 | 5 | yes | occasionally | sometimes | high |
| 950 | 60 or older | Male | no | yes | less than half an hr | 27.0 | no | yes | 6 | 5 | yes | occasionally | very often | high |

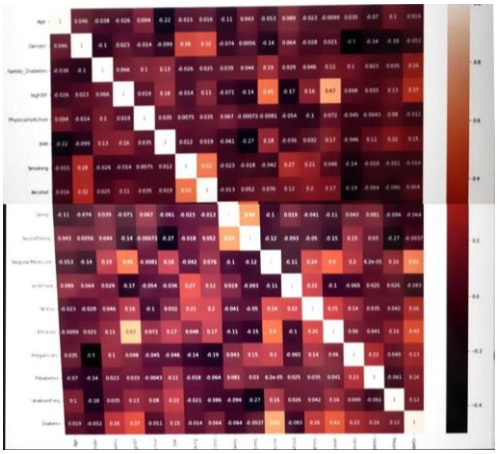


| PhysicallyActive | BMI | Smoking | Alcohol | Sleep | SoundSleep | RegularMedicine | JunkFood | Stress | BPLevel | Pregancies | Pdiabetes | UriationFreq | Diabetic |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| one hr or more | 39.0 | no | no | 8 | 6 | no | occasionally | sometimes | high | 0.0 | 0 | not much | no |
| less than half an hr | 28.0 | no | no | 8 | 6 | yes | very often | sometimes | normal | 0.0 | 0 | not much | no |
| one hr or more | 24.0 | no | no | 6 | 6 | no | occasionally | sometimes | normal | 0.0 | 0 | not much | no |
| one hr or more | 23.0 | no | no | 8 | 6 | no | occasionally | sometimes | normal | 0.0 | 0 | not much | no |
| less than half an hr | 27.0 | no | no | 8 | 8 | no | occasionally | sometimes | normal | 0.0 | 0 | not much | no |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| more than half an hr | 25.0 | no | no | 8 | 6 | no | often | sometimes | normal | 0.0 | 0 | not much | yes |
| more than half an hr | 27.0 | no | no | 6 | 5 | yes | occasionally | sometimes | high | 0.0 | 0 | quite often | yes |
| none | 23.0 | no | no | 6 | 5 | yes | occasionally | sometimes | high | 0.0 | 0 | not much | no |
| less than half an hr | 27.0 | no | yes | 6 | 5 | yes | occasionally | very often | high | 0.0 | 0 | not much | no |

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 905 entries, 0 to 951
Data columns (total 18 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Age                905 non-null    object
 1   Gender             905 non-null    object
 2   Family_Diabetes    905 non-null    object
 3   highBP             905 non-null    object
 4   PhysicallyActive   905 non-null    object
 5   BMI                905 non-null    float64
 6   Smoking            905 non-null    object
 7   Alcohol            905 non-null    object
 8   Sleep              905 non-null    int64
 9   SoundSleep         905 non-null    int64
 10  RegularMedicine    905 non-null    object
 11  JunkFood           905 non-null    object
 12  Stress             905 non-null    object
 13  BPLevel            905 non-null    object
 14  Pregancies         905 non-null    float64
 15  Pdiabetes          905 non-null    object
 16  UriationFreq       905 non-null    object
 17  Diabetic           905 non-null    object
```

- The dataset has no null values

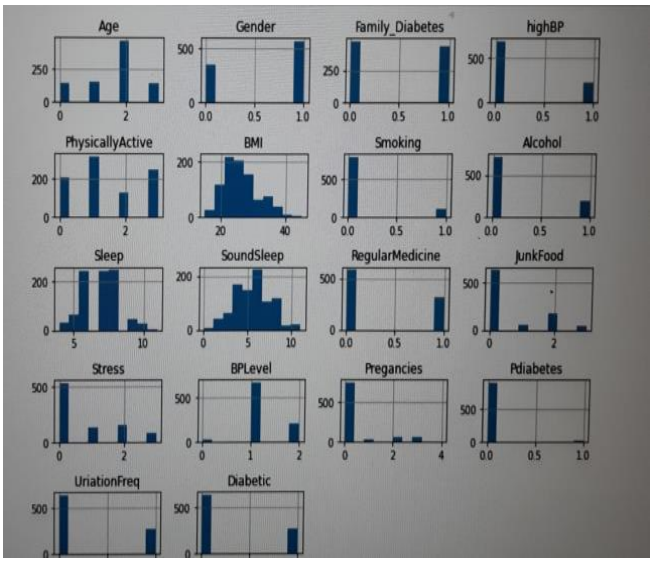
## Heatmap



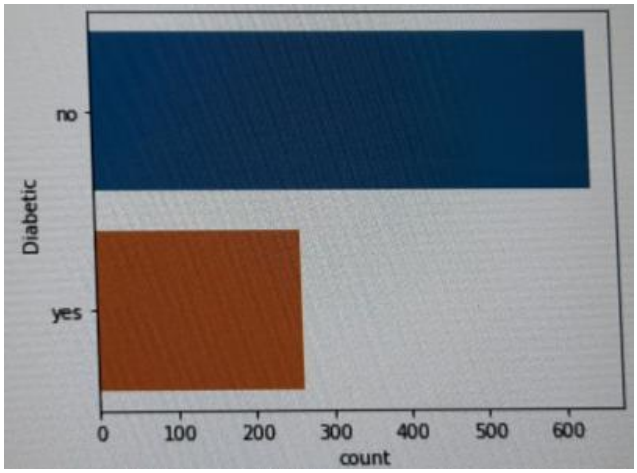
- It's clear that no single feature has a particularly high link with our outcome value. Some characteristics have a negative association with the outcome value, whereas others have a positive correlation.

## Histogram:

Let's look at the plots now. It displays how each feature and label is spread throughout several ranges, demonstrating the importance of scalability. The presence of discrete bars indicates that each of these variables is a category variable. Before using Machine Learning, we'll have to deal with these categorical variables. We have two types of result labels: 0 for no diabetes and 1 for diabetes.




## Bar graph for outcome class

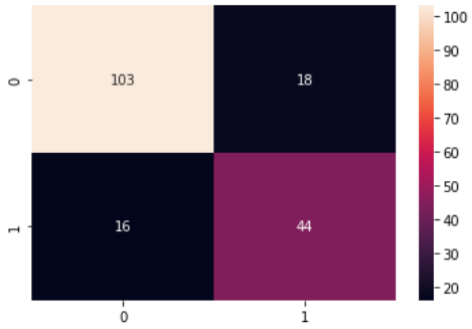



- The graph above illustrates that the data is skewed toward datapoints with a 0 outcome value, indicating that diabetes was not present. Non-diabetic patients outnumber diabetic ones by nearly two to one.

In order to predicting the diabetes, we use three algorithms, which algorithm gives highest accuracy. For finding the accuracy, we take confusion matrix. It will be considered for diabetes prediction.

## Logistic Regression

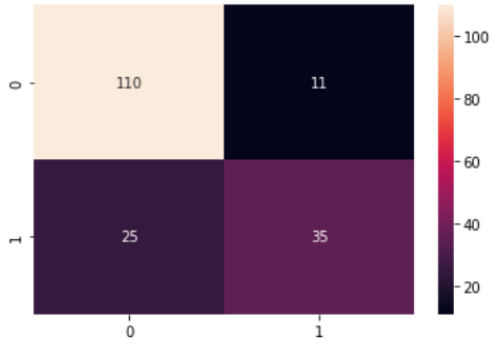
The sigmoid function can be used to classify the output that is a dependent feature, and it employs the probability for classification of the dependent feature in logistic regression.




- Accuracy score of Logistic Regression is 80.11%.

## Support Vector Machine

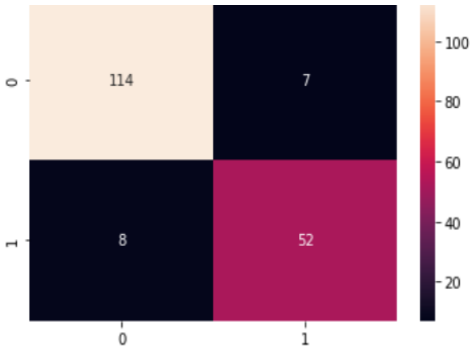
Support vector machine is one of the most widely used supervised learning algorithms, with applications in classification and regression. The algorithm plots each piece of data as a point in n-dimensional space, with the feature value representing the values of each co-ordinate.




- Accuracy score of Support Vector Machine is 81.21%.

## Random Forest

Are an ensemble learning method for classification, regression, and other tasks that works by building a large number of decision trees during training and then outputting the class that is the mode of the classes or the mean prediction of individual trees. Tin Kam Ho used the random subspace approach to construct the first algorithm for random decision forests. It is determined that in order to improve accuracy, it should over train in areas where sensitive data features can be restricted at random.




- Accuracy score of Random Forest is 92 .09%.

## Accuracy score

| Algorithms | Training Accuracy |
|---|---|
| Logistic Regression | 80.11% |
| Support Vector Machine | 81.21%. |
| Random Forest | 92 .09%. |

Table 1

- By using Random Forest algorithm, we predict the diabetes .
- After removing unnecessary attributes, we only used 5 out of the 17 features in the dataset to predict diabetes.
- In this, the diabetic class gives no(0) for no diabetes, yes(1) for diabetes

| Age | highBP | BMI | RegularM | BPLevel | D |
|---|---|---|---|---|---|
| 50-59 | yes | 39 | no | high | n |
| 50-59 | no | 38 | yes | normal | y |
| 40-49 | no | 24 | no | normal | n |
| 60 or older | yes | 26 | yes | high | y |
| 40-49 | no | 27 | no | normal | n |
| 40-49 | yes | 21 | no | high | y |
| less than 40 | no | 24 | no | normal | n |
| less than 40 | no | 20 | yes | low | n |
| 40-49 | no | 23 | no | normal | n |

## VI. CONCLUSION

Various machine learning algorithms are applied to the dataset in this study, and classification is done using various algorithms, with Random Forest providing the highest accuracy. We've seen how machine learning algorithm accuracies compare to datasets. This research could be expanded to see how likely non-diabetic people are to develop diabetes in the coming years.

## VII. FUTURE SCOPE

As a result, we use common classification and ensemble algorithms on the dataset to make predictions. Diabetes is a prevalent chronic disease that can be extremely dangerous to one's health. Diabetes is diagnosed when blood glucose levels are greater than normal, which is caused by insulin secretion or biological factors. Diabetes can harm our bodies in a variety of ways, including causing tissue, kidney, eye, and blood artery dysfunction. Based on physical examination data and consultation with doctors, machine learning may make a preliminary diagnosis of diabetes mellitus. Many techniques, including machine learning methods like Random Forest, Support Vector Machine, Decision Tree, and others, have recently been utilized to predict diabetes.

## VIII. REFERENCES

[1] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference on IˇSMAC,978-1-5090-3243-3,2017.

[2] Ayush Anand and Divya Shakti," Prediction of Diabetes Based on Personal Lifestyle Indicators", 1st International Conference on Next Generation Computing Technologies, 978-1-4673-6809-4, September 2015.

[3] B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 978-1- 5386- 2745-7,2017.

[4] Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing,2015.

[5] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly," Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.

[6] K Sharmila, S Manickam, Efficient Prediction and Classification of Diabetic Patients from big data using R International Journal of Advanced Engineering Research and Science, volume 2 Posted: 2015-09.

[7] G Hari Sassanian, Sekaran, Big Data Analytics Predicting Risk of Readmissions of Diabetic Patients International Journal of Science and Research, volume 4 Posted: 2015-04.

[8] Mamykina L, Heitkemper EM, Smaldone AM, Kukafka R, Cole-Lewis HJ, Davidson PG, Hripcsak G (2017) Personal discovery in diabetes self-management: discovering cause and effect using self-monitoring data. J Biomed Inform 76:1–8.

[9] Soumya D, Srilatha B (2011) Late stage complications of diabetes and insulin resistance. J Diabetes Metab 2(9):1000167.

[10] Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, Shaw JE (2019) Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. Diabetes Res Clin Proact 157:107843.

[11] Papatheodorou K, Banach M, Edmonds M, Papanas N, Papazoglou D (2015) Complications of diabetes.

[12] Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, Shaw JE (2019) Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. Diabetes Res Clin Pract 157:107843.

[13] Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., and Sakr, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: the henry ford exercise testing (FIT) project. *PLoS One* 12:e0179805. doi: 10.1371/journal.pone.0179805.

[14] Habibi, S., Ahmadi, M., and Alizadeh, S. (2015). Type 2 diabetes mellitus screening and risk factors using decision tree: results of data mining. *Glob. J. Health Sci.* 7, 304–310. doi: 10.5539/gjhs.v7n5p304.