DATA SCIENCE PROJECT REPORT

On

**DIABETES DATASET EXPLORATION USING DIFFERENT MACHINE LEARNING TECHNIQUES**

Submitted as partial fulfilment for the Award of the

DATA SCIENCE CERTIFICATION

By

**Shobhit Sinha**

**(2000320100160.)**

**Saransh Chauhan**

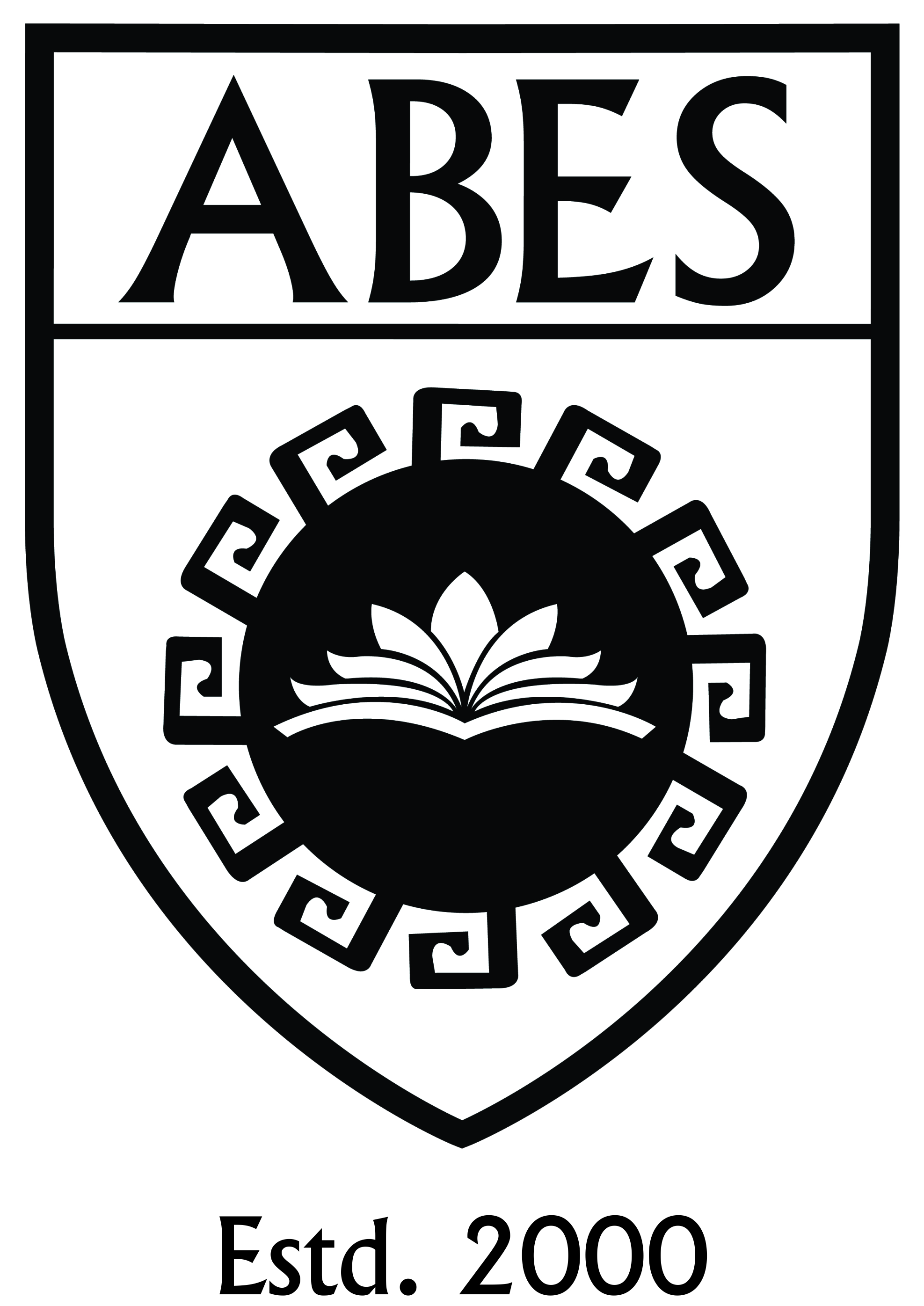
**(2000320100146)**

**Sawan**

**(2000320100153)**

Under the Supervision of

**Ms. SHANU SHARMA**



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

ABES ENGINEERING COLLEGE, GHAZIABAD

AFFILIATED TO

DR. A.P.J ABDUL KALAM TECHNICAL UNIVERSITY, U.P., LUCKNOW

INDIA, 2023

**DECLARATION**

We here by declare that the work being presented in this report entitled “DIABETES DATASET EXPLORATION USING DIFFERENT MACHINE LEARNING TECHNIQUES” is an authentic record of our own work carried out under the supervision of Ms. Shanu Sharma.

**Shobhit Sinha**

**(2000320100160)**

**Saransh Chauhan**

**(2000320100146)**

**Sawan**

**(2000320100153)**

**Date: 19/05/23**

**CERTIFICATE**

This is to certify that Project Report entitled “Diabetes Dataset exploration using different Machine Learning Techniques” which is submitted by Shobhit Sinha, Saransh Chauhan, Sawan in partial fulfillment of the requirement for the Data Science Certification in Department of Computer Science & Engineering, ABES Engineering College (Affiliated to A.K.T.U, Lucknow) is a record of the candidate own work carried out by him under my supervision.

Shanu Sharma

Assistant Professor

Department of Computer Science & Engineering

ABES Engineering College, Ghaziabad, India

Date: 19/05/23

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| Topic | Page No. |
| **INTRODUCTION** | 9-10 |
| **LITERATURE STUDY** | 11-11 |
| **PROPOSED APPROACH** | 12-14 |
| **IMPLEMENTATION AND RESULTS** | 15-15 |
| **CONCLUSION** | 16-16 |
| **REFERENCES** | 17-17 |

**LIST OF FIGURES**

|  |  |
| --- | --- |
| **Figure Description** | **Page No.** |
| Figure 1.2. Training set & test set | 14 |
| Figure 1.1. Comparison of result using KNN, SVM, Decision tree | 15 |
|  |  |
|  |  |

**LIST OF TABLES**

|  |  |
| --- | --- |
| **Table Description** | **Page No.** |
| Table 1.1. Missing values | 12 |
| Table 1.2. Original data v/s normalised data | 14 |
| Table 1.3 Comparison of Result of different algorithm | 15 |
|  |  |

**ABSTRACT**

Global health healthcare might be brought on by the chronic illness of diabetes. The International Diabetes Federation estimates that 382 million people worldwide have diabetes. This will increase to 592 million by the year 2035. The condition known as diabetes mellitus, or just diabetes, is brought on by an elevated blood glucose level. Diabetes may be diagnosed using a number severe national techniques based on physical and chemical examinations. Although diabetes affects human organs including the kidney, eye, heart, nerves, foot, and others, early diabetes prediction is a difficult task for medical professionals owing to intricate dependencies on many elements. By providing fresh light on familiar topics, data science techniques have the potential to advance other scientific disciplines. Making forecasts using medical data is one of these tasks. In the realm of data science, machine learning is a young scientific discipline that examines how machines learn via experience. The goal of this research is to create a system that, by integrating the outcomes of several machine learning approaches, can accurately predict diabetes in a patient early on. Through the use of three distinct supervised machine learning techniques, including SVM, Logistic Regression, and KNN, this study tries to predict diabetes. Additionally, this project seeks to present a practical method for combining flask for end-to-end deployment and machine learning techniques to detect diabetic illness sooner.

**CHAPTER 1**

**INTRODUCTION**

* 1. **Problem Statement**

There are various ongoing illnesses everywhere, both in developed and underdeveloped countries. Diabetes is one of these illnesses. Diabetes is a metabolic disorder that raises blood sugar levels by generating either a large amount of insulin or a small amount of insulin in the body. Perhaps the worst disease on the globe is diabetes. It is not just a disease but also a contributor to other illnesses including heart failure, blindness, renal problems, nerve damage, etc. Therefore, early detection of such a chronic metabolic disorder might assist medical professionals all around the world in preventing the loss of human life. Currently, with the rise of machine learning, AI, and neural networks, as well as their use in other fields [1, 2], we may have the possibility of discovering a solution to this problem. Scientists can discover new realities from current informational indexes connected to wellbeing using ML techniques and neural systems, which may aid in monitoring and detecting illnesses. The Diabetes Database of the Pima Indians was used to finish the current project. The goal of this approach is to create a machine learning model that can accurately predict the chance or probabilities of a patient having diabetes. The patient must go to a symptomatic centre as part of the standard procedure for diagnosing diabetes. Getting exact results from the information is one of the main challenges in bioinformatics analysis. The process of illness identification may become complicated by human error or by multiple scientific testing. This model can determine if a patient has diabetes or not, assisting professionals in ensuring that the patient in need of clinical attention may receive it on time and also assisting in preventing the loss of human lives.

* 1. **Motivation**

Diabetes is a chronic disease that can pose a global public health crisis. Diabetes mellitus, or diabetes, is a disease caused by elevated blood sugar levels. It is one of the deadliest diseases on the rise worldwide.

Various traditional methods based on physical and chemical examination are available for diagnosing diabetes. Data science techniques have the potential to benefit other scientific disciplines by shedding new light on frequently asked questions. Machine learning is a new branch of data science that deals with how machines learn from experience.

This study focuses on recent developments in machine learning that have had a significant impact on diabetes detection and diagnosis. We also aim to propose an effective method for early detection of diabetes.

* 1. **Objective**

i) The primary aim of this project is to analyse the Diabetes Dataset and use Support Vector Machine, K-Nearest Neighbors & Decision Tree algorithms for prediction.

ii)Allow users to predict diabetes utilizing the prediction engine.

iii)The objective is set to achieve the aims of the project through a Research on statistical models in machine learning and to understand how the algorithms works

* 1. **Scope of Work**

Our study could be extended to examine the likelihood that non-diabetics will develop diabetes in the next few years.

**CHPATER 2**

**LITERATURE STUDY**

The classification method of diabetes prediction has two mutually exclusive outcomes: either the subject has diabetes or not. After conducting a thorough investigation, we came to the conclusion that while a variety of classification algorithms may be employed for prediction, the observed accuracy varied. When applied to our dataset, the methods logistic regression, KNN, Naive Bayes [3, random forest, decision tree, and neural network [4] performed comparably to those found in popular publications. The accuracy rate for KNN and logistic regression approaches was 80%.

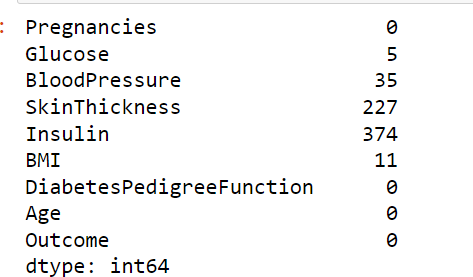
Our algorithm's decision was primarily impacted by its flexibility and compatibility with upcoming applications. Neural networks are the obvious choice as data storage because of the inevitable transition toward DNA. Neurons are used by neural networks to convey information between multiple layers, with each node focusing on a distinct weighted parameter to aid in the prediction of diabetes.

The goal of this framework is to create a machine learning (ML) model that can predict with accuracy the likelihood or odds of a patient having diabetes. The patient must attend an asymptomatic center of the standard procedure for diagnosing diabetes. Getting exact results from the information is one of the main challenges in bioinformatics analysis. The process of identifying the disease might become complicated by human error or by multiple scientific testing. This model can predict if a patient has diabetes, assisting professionals in ensuring that patients in need of therapeutic attention can receive it on time and also assisting in predicting the loss of human lives.

**CHPATER 3**

**PROPOSED APPROACH**

1. **Dataset collection** - This involves gathering data, interpreting it to look for hidden patterns and trends that may be used to make predictions, and assessing the outcomes. The dataset has 10 columns, or the total number of features, and 1405 rows, or the total amount of data. Features comprise Age, Diabetes PedigreeFunction, Skin Thickness, Insulin, Glucose, Blood Pressure, and Pregnancies.
2. **Data Pre-processi**ng: This stage of the model deals with erroneous data to provide more precise and accurate results. For example, the Id field in this dataset is inconsistent, thus we eliminated the feature. Missing values are not present in this dataset. Due to the fact that certain variables, such as age, skin thickness, blood pressure, and BMI, cannot have zero values, we imputed missing values for these attributes. Then, using StandardScaler, data was scaled. No feature selection was done because there weren't many features that were both useful for prediction and little in quantity.
3. **Missing value identification**: Table 2 shows how we determined the missing values in the datasets using the Panda library and SK-learn. The appropriate mean value was used to fill in the missing value.



Missing value identified

### Feature selection:

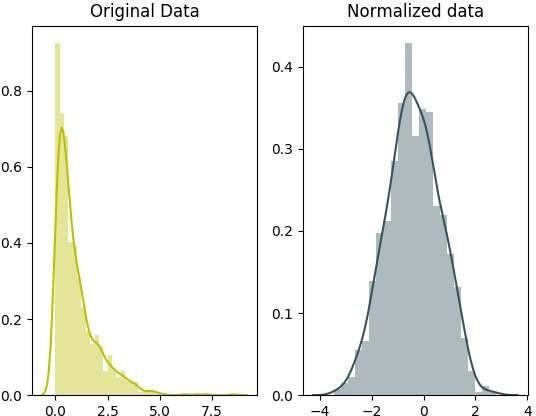
It's common practise to use Pearson's correlation approach to identify the most pertinent characteristics or qualities. This approach calculates the correlation coefficient, which links the input and output qualities. The coefficient value stays within the range of 1 to 1. A significant correlation is indicated by a value above or below 0.5, and by a value of zero, there is no correlation.

|  |  |
| --- | --- |
| Attributes | Correlation coefficient |
| Glucose | 0.484 |
| BMI | 0.316 |
| Insulin | 0.261 |
| Preg | 0.226 |
| Age | 0.224 |
| Skin Thickness | 0.193 |
| BP | 0.183 |
| DPF | 0.178 |

### Scaling and Normalization:

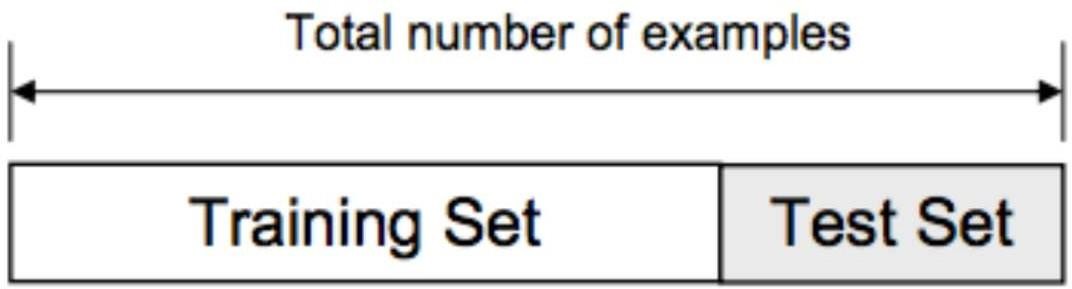
To speed up the calculation of the method, we performed feature scaling by normalising the data in the 0 to 1 range.

Scaling is the process of changing data to match a predetermined range, such as 0-100 or 0-1. When employing techniques based on estimates of how far away data points are, such as support vector machines (SVM) or k-nearest neighbors, you should scale the data (KNN). These algorithms assign the same emphasis to changes of "1" in any numerical characteristic.



### Splitting of data:

The dataset is prepared for training and testing after data cleaning and preprocessing. We randomly divide the dataset into the training and testing sets when using the train/split approach. We collected 1600 samples for training, and 400 samples for testing.



1. **Design and implementation of the classification model**: In this study, in- depth analyses are carried out using a variety of ML classification approaches, including DT, KNN, RF, NB, LR, and SVM.
2. **Machine learning classifier:** Using machine learning techniques, we created a model. to predict diabetes, several classifiers and ensemble approaches were used. To analyze the performance by determining the accuracy of each classifier, we used SVM, LR, DT, and RF machine learning classifiers. Python's sci-kit-learn packages are used to implement each classifier. The next section provides a description of the used classification algorithms

**CHPATER 4**

**IMPLEMENTATION AND RESULTS**

Algorithms for machine learning categorization were created to predict diabetes at an earlier stage. 30% of the data were utilised for testing, and 70% were used for trining. We discovered that the Random Forest Classifier predicted with 79% accuracy in this data splitting ratio, making it the dataset's greatest accuracy. Below is a comparison of the output from each classifier that has been used.

|  |  |
| --- | --- |
| **Machine Learning Algorithm** | **Result** |
| ***SVM Test Accuracy*** | ***77*** |
| ***DECISION TREE Test Accuracy*** | ***75*** |
| ***KNN Test Accuracy*** | ***72*** |

Result



**CHPATER 5**

**CONCLUSION**

The project's goal was to explore diabetes dataset using different Machine Learning Algorithms create a model that could recognize diabetic individuals who are most likely to be admitted to the hospital. The challenge of predicting the likelihood of hospital admission is very challenging. This procedure's outcome is influenced by a variety of circumstances. Methods to improve healthcare institutions' understanding of what matters in forecasting the probability of hospital admission are urgently needed right now. By suggesting a system that may be utilized as a helper in identifying the patients at higher risk of having diabetes, this effort makes a tiny addition to the already employed techniques of diabetes detection. This study does this by examining a number of important indicators, including the patient's body mass index and blood glucose level. employing a variety of machine learning algorithms, as well as through a review of patient medical information.

In this study, various machine learning algorithms were applied to the data set to Classification was performed with different algorithms. It is observed that the accuracy achieved through the use of SVM was 77 percent, accuracy using Decision Tree was 75 percent and accuracy achieved through KNN was 72 percent. Of these, SVM has the highest accuracy of 77%. We compared the accuracy of machine learning algorithms using same dataset. It is clear that models using this dataset improve the accuracy and precision of diabetes prediction compared to existing datasets. Additionally, this study could be extended to examine the likelihood that non-diabetics will develop diabetes in the next few years.

**REFERENCES**

* 1. Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance’s for Diabetes Prediction using Machine Learning". IEEE, pp 942- 928, 2018.
  2. K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Compu- tation Automation and Networking, 2019.
  3. Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Perfor- mance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 Feb- ruary, 2019.
  4. An evolving SVM model for DDOS attack detection in software defined networks is presented by Sahoo, K.S., et al. 132502–132513 in IEEE Access 8 (2020)
  5. A machine learning method for forecasting DDoS traffic in software defined networks by Sahoo, K.S., et al. International Conference on Information Technology, 2018, p. (ICIT). IEEE (2018) (2018)
  6. Jakka, A., and Vakula Rani, J.: Performance assessment of machine learning models for diabetes prediction. Int. J. Innov. Technol. Explor. Eng. (IJITEE) 8(11) (2019). ISSN: 2278-3075
  7. Using machine learning approaches, Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., and Tang, H. predicted diabetes mellitus. J. Front. Genet. Bioinform. Comput., published: 06 2018
  8. <https://en.wikipedia.org/wiki/Bin_ary_classification>
  9. <https://www.webmd.com/>
  10. <https://archive.ics.uci.edu/ml/index.php>