Diabetes Prediction Using Machine Learning

Shruthy Radhakrishan

ABSTRACT

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to the International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes is a disease caused due to the increased level of blood glucose. This high blood glucose produces the symptoms of frequent urination, increased thirst, and increased hunger. Diabetes is one of the leading causes of blindness, kidney failure, amputations, heart failure and stroke. When we eat, our body turns food into sugars, or glucose. At that point, our pancreas is supposed to release insulin. Insulin serves as a key to open our cells, to allow the glucose to enter and allow us to use the glucose for energy. But with diabetes, this system does not work. Current practice in hospitals is to collect required information for diabetes diagnosis through various tests and appropriate treatment is provided based on diagnosis. Big Data Analytics plays a significant role in healthcare industries. Healthcare industries have large volume databases. Using big data analytics one can study huge datasets and find hidden information, hidden patterns to discover knowledge from the data and predict outcomes accordingly. In the existing method, the classification and prediction accuracy is not so high. In this paper, we have proposed a diabetes prediction model for better classification of diabetes which includes factors like Glucose, BMI, Age, Insulin, etc. Classification accuracy is boosted with the new dataset compared to the existing dataset. Further with imposed a pipeline model for diabetes prediction intended towards improving the accuracy of classification.

Keywords:Machine Learning, Diabetes, Decision tree, K nearest neighbour, Logistic Regression, Support vector Machine, Accuracy.

1. INTRODUCTION

Diabetes Is the fastest growing disease among the people even among the youngsters. In understanding diabetes and how it develops, we need to understand what happens in the body without diabetes. Sugar (glucose) comes from the foods that we eat, specifically carbohydrate foods. Carbohydrate foods provide our body with its main energy source everybody, even those people with diabetes, needs carbohydrates. Carbohydrate foods include bread, cereal, pasta, rice, fruit, dairy products and vegetables (especially starchy vegetables). When we eat these foods, the body breaks them down into glucose. The glucose moves around the body in the bloodstream. Some of the glucose is taken to our brain to help us think clearly and function. The remainder of the glucose is taken to the cells of our body for energy and also to our liver, where it is stored as energy that is used later by the body. In order for the body to use glucose for energy, insulin is required. Insulin is a hormone that is produced by the beta cells in the pancreas. If the pancreas is not able to produce enough insulin (insulin deficiency) or if the body cannot use the insulin it produces (insulin resistance), glucose builds up in the bloodstream (hyperglycaemia) and diabetes develops.To reduce the possibility of developing some serious complications related to diabetes, machine learning and data mining techniques can be applied to diabetes-related datasets.

2. PROBLEM STATEMENT

NIDDK (National Institute of Diabetes and Digestive and Kidney Diseases) research creates knowledge about and treatments for the most chronic, costly, and consequential diseases.The dataset used in this project is originally from NIDDK.Doctors rely on common knowledge for treatment. When common knowledge is lacking, studies are summarized after some number of cases have been studied. But this process takes time, whereas if machine learning is used, the patterns can be identified earlier. The objective is to predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.

3. CUSTOMER NEED ASSESSMENT

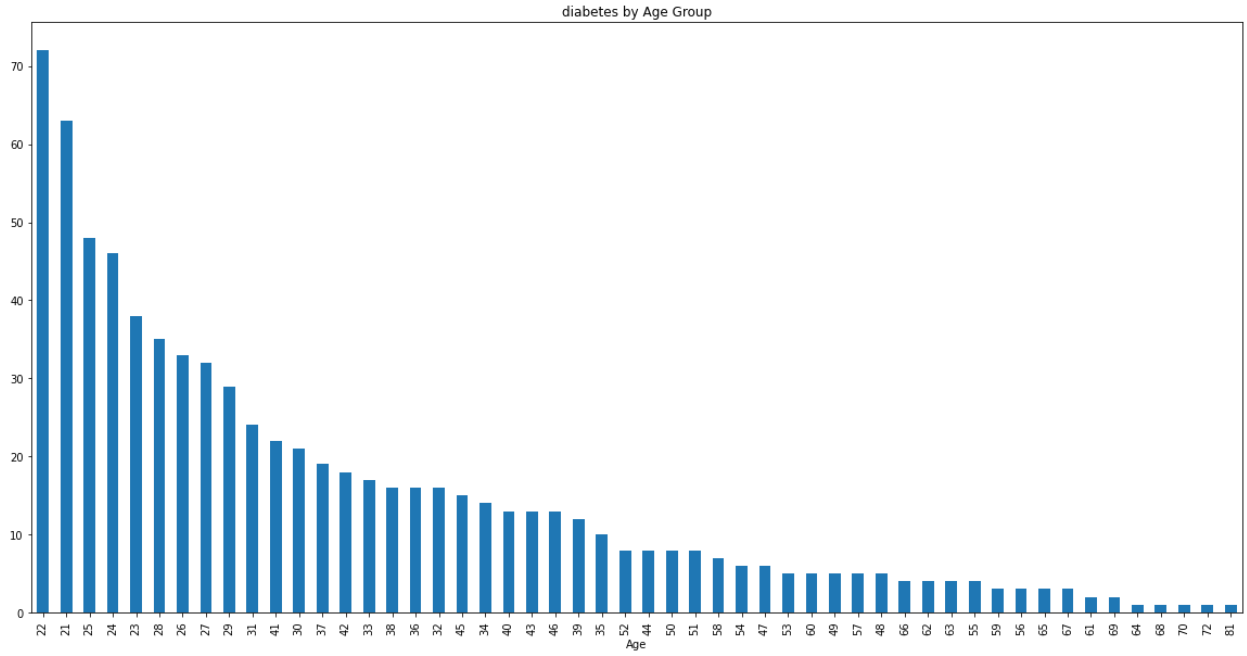
According to the CDC's 2017 National Diabetes Statistics Report, 30.3 million individuals in the U.S. have diabetes, and 7.2 million people living with the disease are undiagnosed. This showcases the need for diabetes management programs and continuous patient education and monitoring. According to a study published in Diabetes Spectrum, pharmacist-led interventions in a rural primary care clinic were associated with the majority of patients experiencing an A1C reduction of at least 1%, which has the potential to reduce the risk of complications and decrease diabetes associated costs. Pharmacists' have a unique skillset as drug experts and play an integral role as part of the healthcare team-not only improving patient outcomes, but also expanding business opportunities.

4. EXTERNAL SEARCH

- https://www.who.int/news-room/fact-sheets/detail/diabetes
- https://www.healthline.com/health/diabetes
- The dataset which is used was taken from the kaggle :-

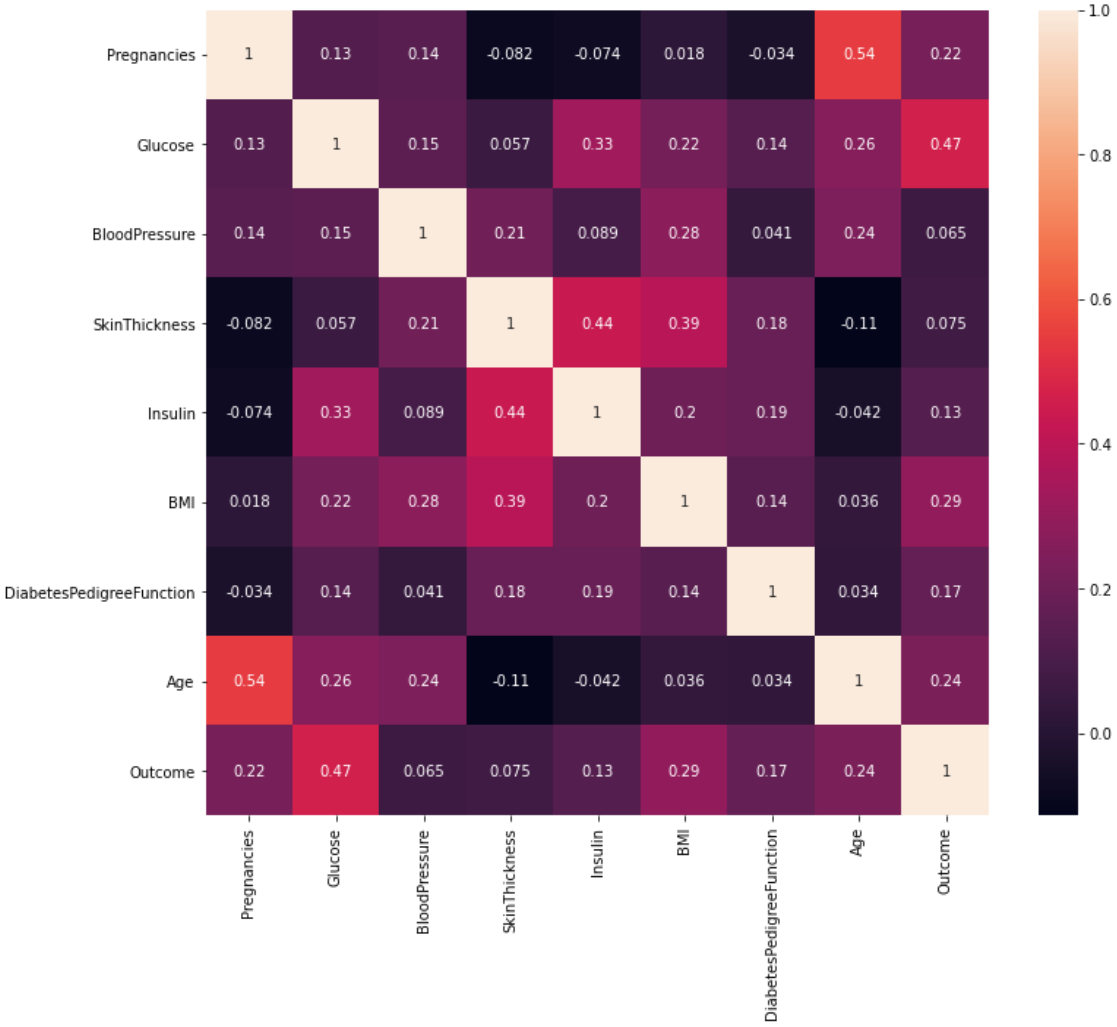
  https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

5. TARGET SPECIFICATIONS AND CHARACTERIZATION



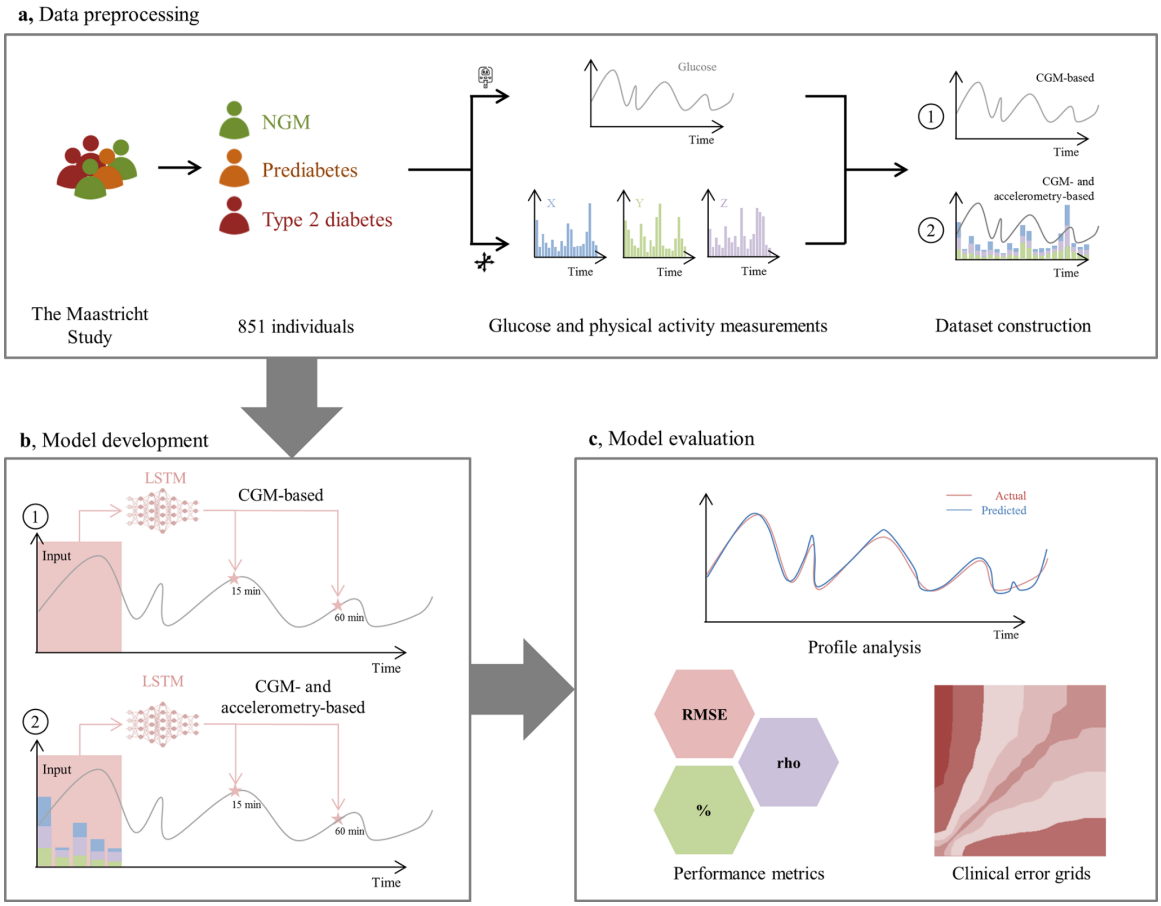
In the above figure we can see that age between 21-35 have been associated with risk of diabetes that of patients of older age. More than 9 out of 10 people with diabetes have type 2. It used to be called adult-onset diabetes because it was rarely diagnosed in children. Age is a big risk factor for type 2. The older you are, the more likely you are to have it. That also holds true for preteens and teenagers, whose diabetes rates have climbed sharply in recent years. You can have diabetes for years and not know it. Symptoms like thirst, peeing more often, blurry eyesight, and tingling hands and feet may come on slowly without your noticing.

6. BENCHMARKING



7. APPLICABLE PATENTS

Closed-loop insulin delivery systems, which integrate continuous glucose monitoring (CGM) and algorithms that continuously guide insulin dosing, have been shown to improve glycaemic control. The ability to predict future glucose values can further optimize such devices. In this study, we used machine learning to train models in predicting future glucose levels based on prior CGM and accelerometry data. Machine learning-based models are able to accurately and safely predict glucose values at 15- and 60-minute intervals based on CGM data only. Future research should further optimize the models for implementation in closed-loop insulin delivery systems.

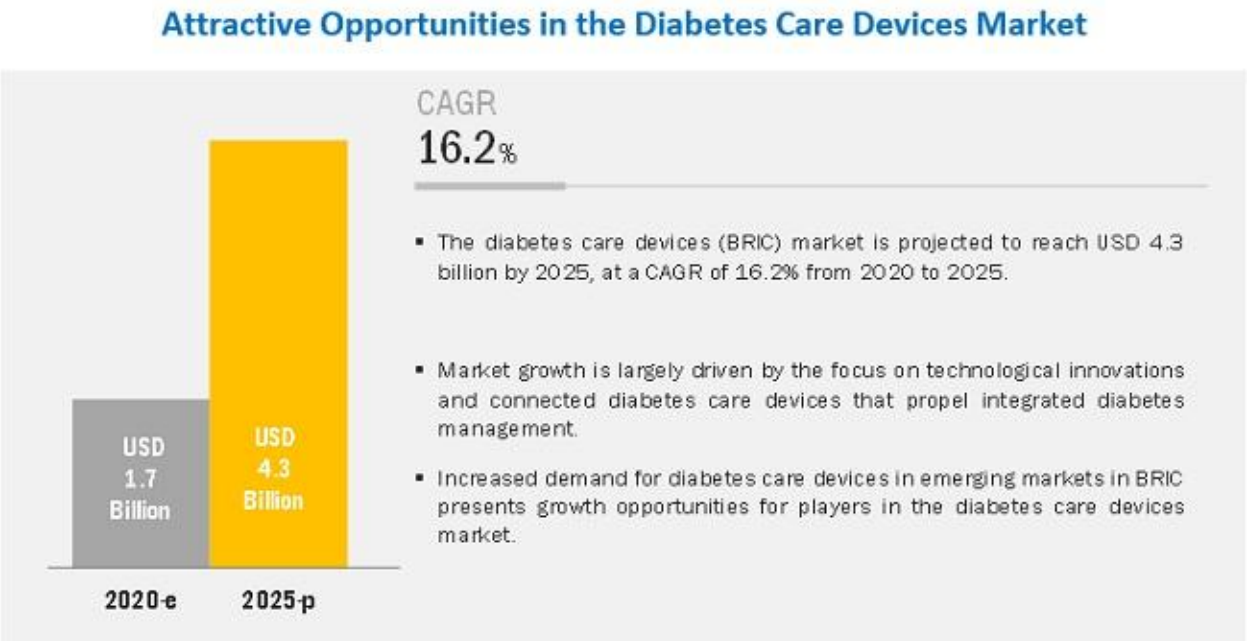


8. Applicable Constraints

1) Need a lot of space to store the data gathered over the net.
2) Pandas function is utilized to read CSV.file where the data set file is in excel format.
3) Used matplotlib to visualize data for better understanding purposes.
4) For modeling different types of classification algorithms are applied.

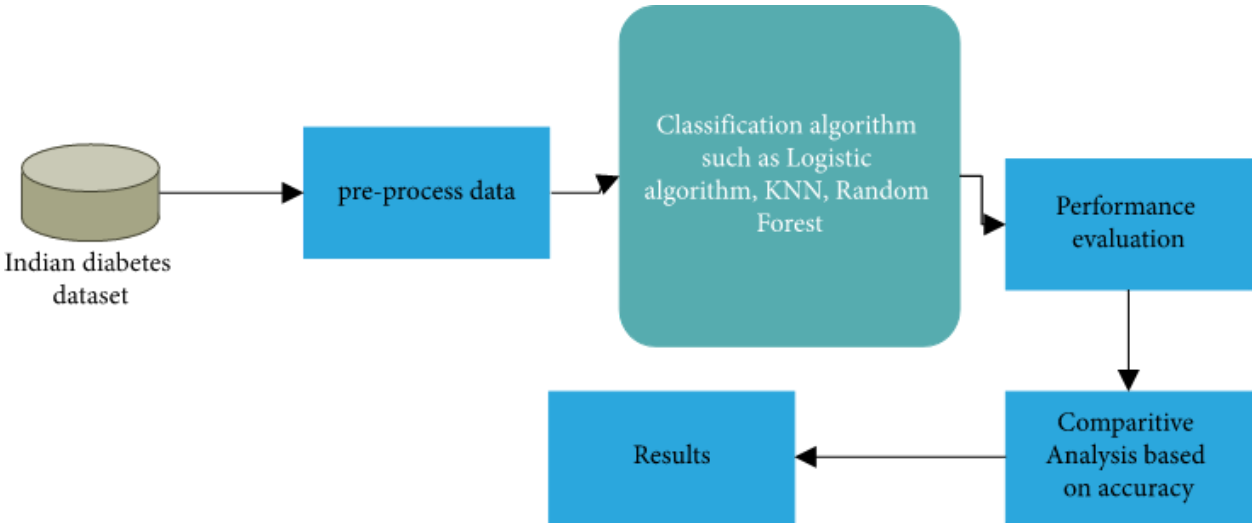
9. Business Opportunity

The diabetes care devices (BRIC) market size is projected to reach USD 4.3 billion by 2025 from USD 1.7 billion in 2020, at a CAGR of 16.2% during the forecast period. Factors such as the increasing diabetic population, the rising awareness of diabetes treatment and management, and favorable national health strategies are expected to drive the growth of the diabetes care devices (BRIC) market.



10. Final product Prototype

The proposed framework is divided into different phases. The flow diagram is illustrated in. Python Jupyter Note was used for the entire implementation. Different packages such as NumPy, pandas, scikit, and Matplotlib have been used in analyzing the data. The task performed in each phase and the relevant functions explored from Python tool kits are described below.



10.1 Data Set

Pima Indian Diabetes Database is a familiar and commonly used data set for the prediction of diabetes. This data set consists of 768 rows and 9 columns. The attributes included in the column are glucose, pregnancies, skin thickness, blood pressure, BMI, insulin, age, and outcomes. The outcome variable predicts whether the patient is diabetic positive or diabetic-negative.

```
In [2]: diabetes_df = pd.read_csv('health care diabetes.csv')
        diabetes_df
```

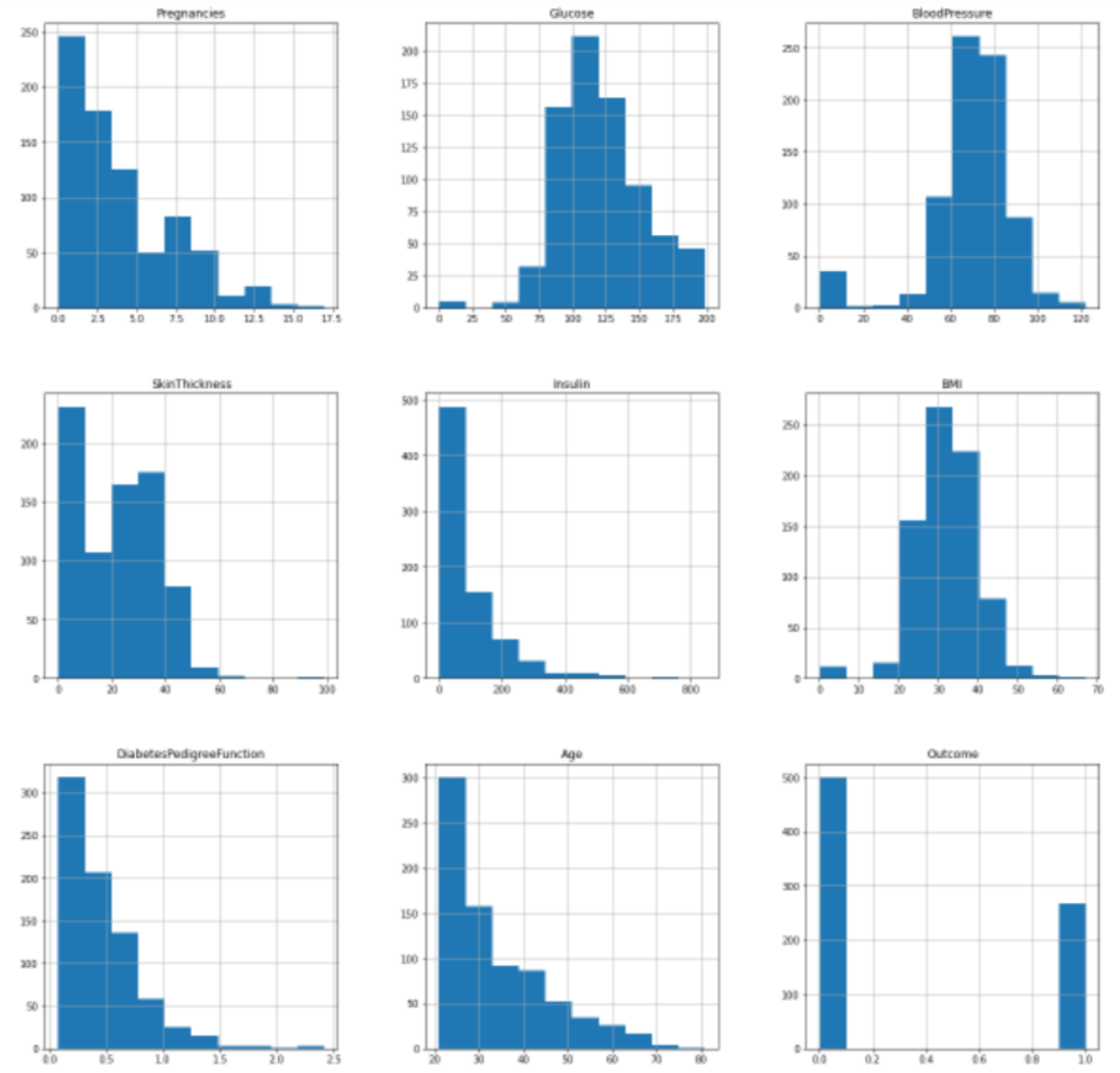
Out[2]:

	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

10.2 Data Visualization

Data visualization helps to understand the data better by putting it in a visual form. In this phase, data are represented in the form of bar chart. The analysis reveals the percentage of people affected by diabetes diseases. It also displays the information of the data set such as age, blood pressure, pregnancies, and glucose. Apart from that, it predicts how many people are affected by diabetes from 768. For displaying output, the graphical representation functions such as plot axis, pyplot, and several others have been used.



10.3 Preprocessing

This section includes the removal of outliers and standardizing the data. The processed data have been used for creating a model. The data should be preprocessed and arranged properly before applying classifiers to the data index. These data should be handled carefully before connecting.

This data set contains missing values. Few selected attributes such as blood pressure, skin

thickness, glucose level, and BMI are assigned with missing values because these parameters cannot have null values. Then, we normalized all values by scaling the data set.

10.4 Machine Learning Classification Algorithms

Subsequently, after preprocessing the data ML classifiers are applied using the scikit-learn Python Toolkit. Scikit is a simple tool kit used to process and analyze the data. These tool kits are used in most of the work. Foremost using a function like the model selection train test split, the data set is divided into the training and testing data sets. Due to the limited data set source, about 90% of the data set is used for training purposes and the remaining 10% is used for testing by selecting the data randomly. Then, different classifiers such as ML algorithms are applied to diagnose diabetes. ML classifiers are adapted because of their simplicity and popularity. Since this work focuses on hyper-parameter tuning, it will be explained in the succeeding section.

10.5 Comparison

In this section, the ML classification algorithm is compared based on accuracy. After the evaluation process, one of the best ML classifiers is identified and hyper-parameter tuning has been applied to produce the best result. ML technique is considered valuable in diagnosing the disease. Early diagnosis advantages the patients with early medical attention. In this study, few existing ML classification models for the prediction of diabetic patients have been discussed based on the accuracy. An expression of accuracy on the classification problem has been identified. ML technique was enforced on the data set. It was trained and confirmed on the test data set and verified. The results of our implementation method show how the Randomforest performed better than other Ml algorithms.

11. CONCLUSION

The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully. The proposed approach uses various classification and ensemble learning method in which SVM, Knn, Random Forest, Decision Tree, Logistic Regression and Gradient Boosting classifiers are used. And 77% classification accuracy has been achieved. The Experimental results can assist health care to take early prediction and make early decision to cure diabetes and save humans life.

12. REFERENCE

- Human Disease Classification and Segmentation using Machine and Deep Learning
- How Age Relates to Type 2 Diabetes
- Diabetes Care Devices Market

GitHub Link :-https://github.com/shruthyR/Diabetes-Prediction.git