**1 INTRODUCTION**  
**1.1  Overview** **:**

Diabetes mellitus is a chronic disease characterized by hyperglycemia. It may cause many complications. According to the growing morbidity in recent years, in 2040, the world’s diabetic patients will reach 642 million, which means that one of the ten adults in the future is suffering from diabetes. There is no doubt that this alarming figure needs great attention.

Diabetes Mellitus (DM) and its complications have become a serious public health problem and are an important cause of morbidity, mortality and disability rates throughout the world. Globally the DM day of the world is celebrated every November 14th, this proves that DM is a global problem that occurs in every country, both in developed countries and in poor and developing countries. Diabetes Mellitus Type 2 (DMT2) is caused by age, environmental conditions, unhealthy lifestyles such as consuming foods or drinks that are high in sugar, smoking and rarely exercising and Diabetes Mellitus Type 1 (DMT1) is caused by genetic factors (hereditary disease inherited from previous offspring) and autoimmune (destruction of pancreatic beta cells that have little or no insulin secretion can be determined by the level of cpeptide protein whose numbers are few or not detected at all). In addition, DMT1 and DMT2 need to do early detection because DM is a disease that tends to be less aware, usually someone only realizes it if they already experience complications of DM.

With the rapid development of machine learning, machine learning has been applied to many aspects of medical health for accurate predictions.  
**1.2  Purpose :**

The purpose of the project is to build a machine learning model that can efficiently discover the rules to predict the risk level of patients based on the given parameter about their health (In this project, we can use decision tree, random forest and neural network to predict diabetes mellitus). Then we evaluate the performance of the model in terms of different parameter like classification accuracy AUC-ROC Curves.  
An android application is built from where the patient health features are entered and depending on the entered parameters, the machine learning model integrated to application will predict the type of diabetes and according to the type of diabetes diet plan for the person will be displayed on the UI.

**2 LITERATURE SURVEY**  
  **2.1  Existing problem :**

The existing problem involves taking medical tests by the patient which may take time, not beneficial for time critical circumstances. Or, considering that the details of a patient is available, the doctor has to make an educated guess to predict the disease which might not be efficient or always correct prediction.

**2.2  Proposed solution :**

The proposed solution is to usethe/our machine learning software that predicts the diabetes mellitus disease in a patient with accuracy more than 90 percent (provided with enough training dataset) in fraction of seconds when provided with the data such as age, insulin level, skin thickness, BMI etc.

**3. THEORITICAL ANALYSIS**

  **3.1  Block diagram :**

**3.2  Hardware / Software designing :**

**Hardware Requirements:**

Access to IBM cloud or graphic card and 8gb RAM, intel i5 processor. (specifications are based on the amount of workload )

**Software Requirements:**

Pandas==0.23.4

Matplotlib==3.0.2

Numpy==1.17.3

Scikit\_learn==0.21.3

Flask==1.0.2

Seaborn==0.9.0

Browser==chrome etc

**4 EXPERIMENTAL INVESTIGATIONS**

The following Important steps are followed sequencially for training and evaluation:

1. **DATA COLLECTION**

The dataset used for training the model : <https://www.kaggle.com/uciml/pima-indians-diabetes-database>.

The data is then loaded in the notebook using panda’s dataframe.

**Data Statistical Observation:**

1. There are a total of 768 records and 9 features in the dataset.
2. Each feature can be either of integer or float dataype.
3. Some features like Glucose, Blood pressure , Insulin, BMI have zero values which represent missing data.
4. There are zero NaN values in the dataset.
5. In the outcome column, 1 represents diabetes positive and 0 represents diabetes negative.
6. **DATA PREPROCESSING**

**Library Imports**: Panda, Numpy, Matplotlib, Seaborn, Warnings.

**Data Visualization**: outcome countplot, histogram of ach feature, heatmap, pairplot.

**Observations:**

The countplot tells us that the dataset is imbalanced, as number of patients who don't have diabetes is more than those who do.

From the correaltion heatmap, we can see that there is a high correlation between Outcome and [Glucose,BMI,Age,Insulin]. We can select these features to accept input from the user and predict the outcome.

**Taking care of Missing Data**: The NaN values are replaced with mean values.

**Since all the data is numerical, there is no need of one-hot encoding or label encoding**.

**Feature Scaling**: Minmaxscalar is used to fix the values of features between 0 and 1.

Glucose, Insulin, BMI and Age are selected as the training and prediction features.

**Data Splitting** : The dataset is split as train and test dataset with test size 0.20 (20 %).

1. **MODEL BUILDING**

**Model Training, Testing and evaluation :**

Total of 6 different models are trained with the train dataset and the models are evaluated among them by prediction accuracy for the test dataset.

**The models are :**

Logistic regression

K Nearest Neighbors

Support Vector Machine

Naïve Bayes

Decision Tree

Random Forest

**Observation (Accuracy of each model):**

Logistic Regression: 71.42857142857143

K Nearest neighbors: 78.57142857142857

Support Vector Classifier: 73.37662337662337

Naive Bayes: 71.42857142857143

Decision tree: 68.18181818181817

Random Forest: 75.97402597402598

**(A simple web app is created using python framework FLASK.)**

1. **FLOWCHART**

**6. RESULT**

The users can now use the web interface to get the prediction from the machine learning software.

The K-nearest neighbour gets the highest accuracy: 78.57.

The accuracy depends on the quality and the quantity of the dataset and the design and implementation of the models.

**7 ADVANTAGES & DISADVANTAGES**

The machine learning software is more efficient than the traditional methods of predicting whether a patient has diabetes mellitus. It also offers scaling.

Since the training dataset is comparatively small, the accuracy is less.

**8 APPLICATIONS**

This software can be used in medical profession, by doctors for diabetes mellitus prediction in a patient, besides traditional methods, for more efficiency, scaling and confirmation.

**9 CONCLUSION**

A machine learning model is built that can efficiently discover the rules to predict the risk level of patients based on the given parameter about their health (In this project, we can use decision tree, random forest and neural network to predict diabetes mellitus). Then we evaluated the performance of the model in terms of different parameter like classification accuracy AUC-ROC Curves.  
An android application is built from where the patient health features are entered and depending on the entered parameters, the machine learning model integrated to application is used to predict the type of diabetes and according to the type of diabetes diet plan for the person, displayed on the UI.

**10 FUTURE SCOPE**

With more data, the efficiency and accuracy of the model can be improved.

**11 BIBILOGRAPHY**

[**https://www.kaggle.com/akhilalexander/diabeticprediction**](https://www.kaggle.com/akhilalexander/diabeticprediction)

[**https://towardsdatascience.com/data-visualization-for-machine-learning-and-data-science-a45178970be7**](https://towardsdatascience.com/data-visualization-for-machine-learning-and-data-science-a45178970be7)

[**https://thesmartbridge.com/documents/spsaimldocs/Datapreprocessing.pdf**](https://thesmartbridge.com/documents/spsaimldocs/Datapreprocessing.pdf)

[**https://thesmartbridge.com/documents/spsaimldocs/Datapreprocessing.pdf**](https://thesmartbridge.com/documents/spsaimldocs/Datapreprocessing.pdf)

[**https://thesmartbridge.com/documents/spsaimldocs/Datapreprocessing.pdf**](https://thesmartbridge.com/documents/spsaimldocs/Datapreprocessing.pdf)

[**https://thesmartbridge.com/documents/spsaimldocs/Datapreprocessing.pdf**](https://thesmartbridge.com/documents/spsaimldocs/Datapreprocessing.pdf)

[**https://thesmartbridge.com/documents/spsaimldocs/Datapreprocessing.pdf**](https://thesmartbridge.com/documents/spsaimldocs/Datapreprocessing.pdf)

[**https://thesmartbridge.com/documents/spsaimldocs/Machinelearning.pdf**](https://thesmartbridge.com/documents/spsaimldocs/Machinelearning.pdf)

[**https://www.w3schools.com/bootstrap/bootstrap\_forms\_inputs.asp**](https://www.w3schools.com/bootstrap/bootstrap_forms_inputs.asp)

[**https://thesmartbridge.com/documents/spsaimldocs/FlaskML.pdf**](https://thesmartbridge.com/documents/spsaimldocs/FlaskML.pdf)

**12 APPENDIX**  
**A. Source code :**

**Link for the source code:** [**https://github.com/smartinternz02/llSPS-INT-4160-Diabetes-Milletus-Prediction**](https://github.com/smartinternz02/llSPS-INT-4160-Diabetes-Milletus-Prediction)

**B. UI output Screenshot :**

