UE20CS302: Machine Intelligence
Project Phase – 1

Project Title : Diabetes Prediction using Machine learning
Project ID : 10
Project Team : RITHIKA A PES2UG20CS539
NEERAJA N PES2UG20CS528
SUMEDA PUJA PES2UG20CS562

Abstract and Scope

- Well defined problem statement.
  - Diabetes Mellitus is among critical diseases and lots of people are suffering from this disease.
  - Age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc. can cause Diabetes Mellitus.
  - People having diabetes have high risk of diseases like heart disease, kidney disease, stroke, eye problem, nerve damage, etc.
  - Current practice in hospital is to collect required information for diabetes diagnosis through various tests and appropriate treatment is provided based on diagnosis. This can be extremely inaccurate and not time efficient hence the prominence of machine learning can be seen.

▪ Provide a basic introduction of the project and also an overview of scope it entails.

- Diabetes if not leads to health hazards like: heart related problems, kidney problem, blood pressure, eye damage and can also affects other organs of human body.
- Diabetes can be controlled if it is predicted earlier by using Big Data Analytics whcih plays an 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 existing method, the classification and prediction accuracy is not so high.
- In this project, we have a diabetes prediction model is done for better classification of diabetes which includes few external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc
- To achieve this goal the project work will focus on early prediction of Diabetes in a human body for a higher accuracy through applying various Machine Learning Techniques.

- Major techniques provide better result for prediction by constructing models from datasets collected from patients.
- In this work we will use Machine Learning Classification and ensemble techniques on a dataset to predict diabetes. Logistic Regression (LR) and Support Vector Machine (SVM) algorithm.
- The accuracy is almost the same for both the models.
- The Project work shows that the model is capable of predicting diabetes effectively since the accuracy in high.

Design Approach

What is the design approach followed? And Why?

Support Vector Machine- Support Vector Machine also known as svm is a supervised machine learning algorithm.

Svm creates a hyperplane that separate two classes. It can create a hyperplane or set of hyperplane in high dimensional space. This hyper plane can be used for classification or regression. Svm differentiates instances in specific classes and can also classify the entities which are not sup- ported by data. Separation is done by through hyperplane performs the separation to the closest training point of any class.

Benefits of this approach & are there any drawbacks?

Alternate design approaches, if any.

BENEFITS OF SVM:

1. SVM works relatively well when there is a clear margin of separation between classes.
2. SVM is more effective in high dimensional spaces.
3. SVM is effective in cases where the number of dimensions is greater than the number of samples.
4. SVM is relatively memory efficient.

LIMITATIONS OF SVM:

1. SVM algorithm is not suitable for large data sets.
2. SVM does not perform very well when the data set has more noise i.e. target classes are overlapping.
3. In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform.
4. As the support vector classifier works by putting data points, above and below the classifying hyperplane there is no probabilistic explanation for the classification.

ALTERNATE DESIGN APPROACHES:

1. K-Nearest Neighbour
2. Decision tree

Design Constraints, Assumptions & Dependencies

Discuss the design constraints and assumptions that you have made to select the design approach.

We choose svm method to narrow our data set to get outputs

Infinite lines exist to separate diabetic and non diabetic patients. SVM needs to find the optimal line with the constraint of correctly classifying either class

Follow the constraint: only look into the separate hyperplanes(e.g. separate lines), hyperplanes that classify classes correctly

Conduct optimization: pick up the one that maximizes the margin.

Proposed Methodology / Approach

Details of the new approach- benefits/drawbacks

LOGISTIC REGRESSION

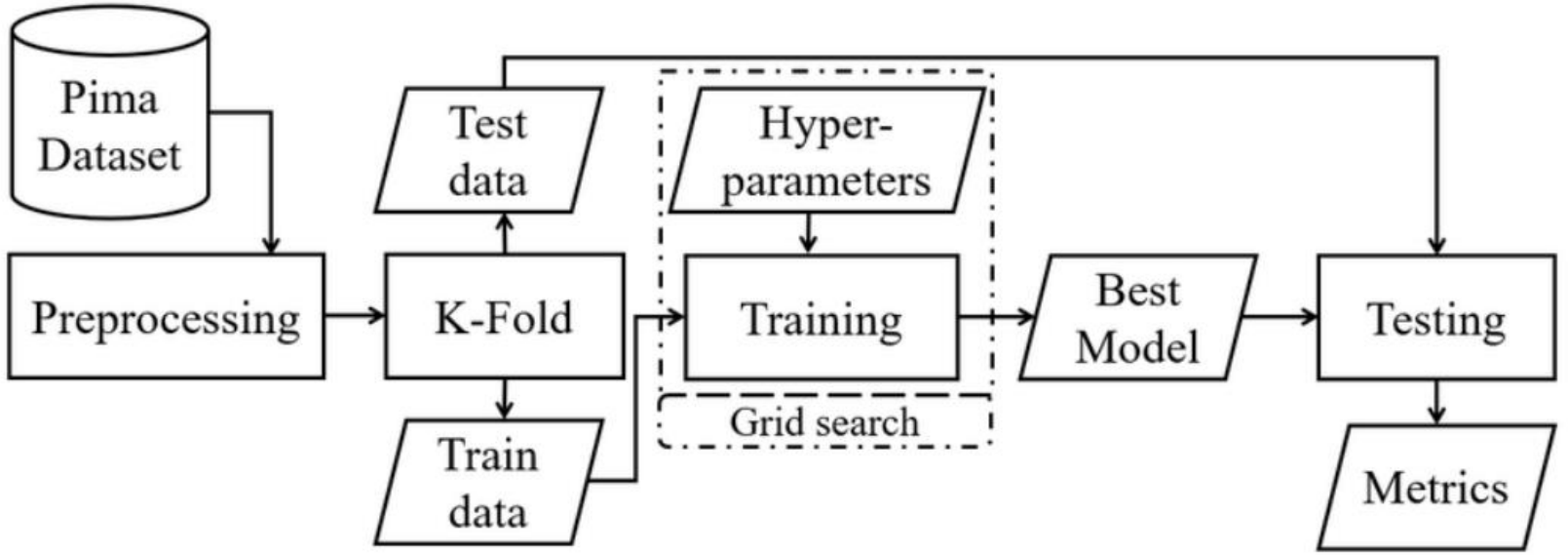
Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Benefits of Logistic Regression :

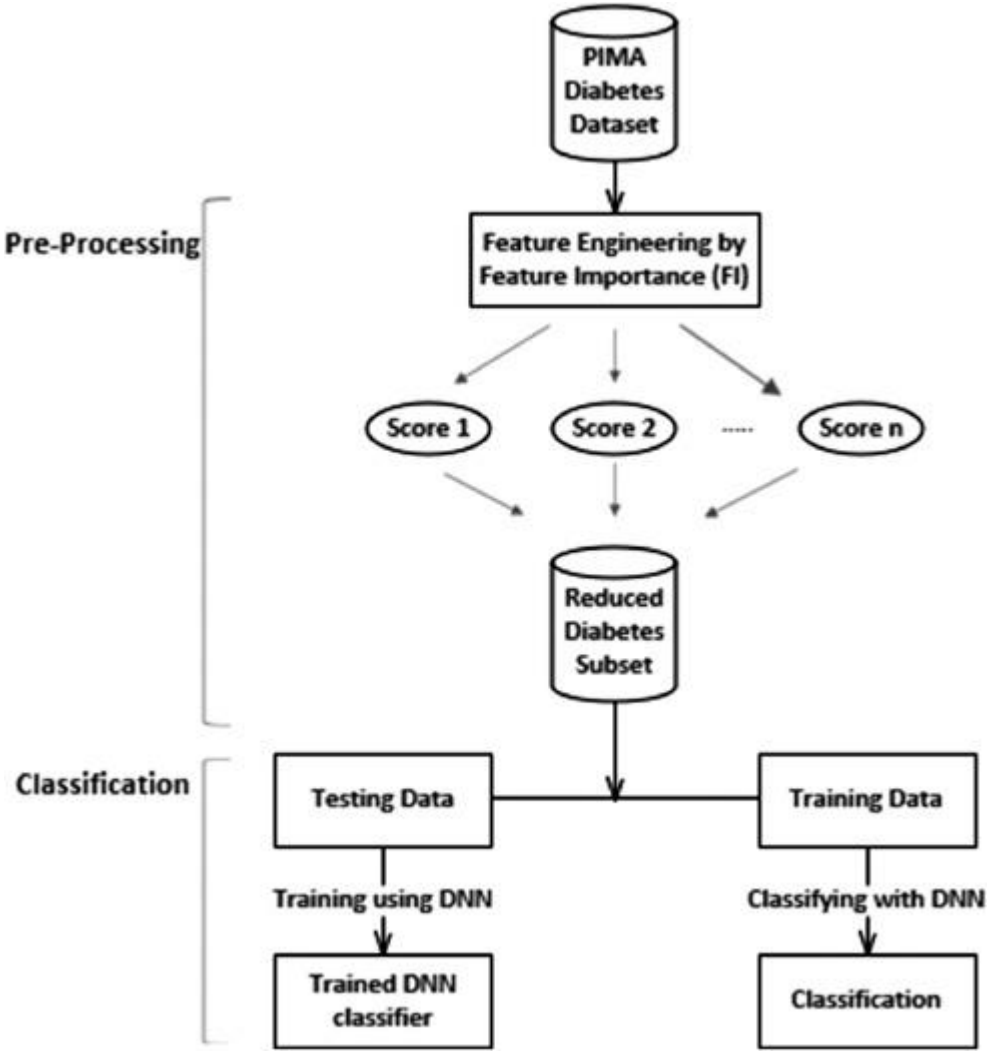
- It is easier to train, implement and predict data.
- Very fast in classifying unknowns
- This algorithm allows models to be updated easily to reflect new data, unlike decision trees or support vector machines. The update can be done using stochastic gradient descent.

Architecture

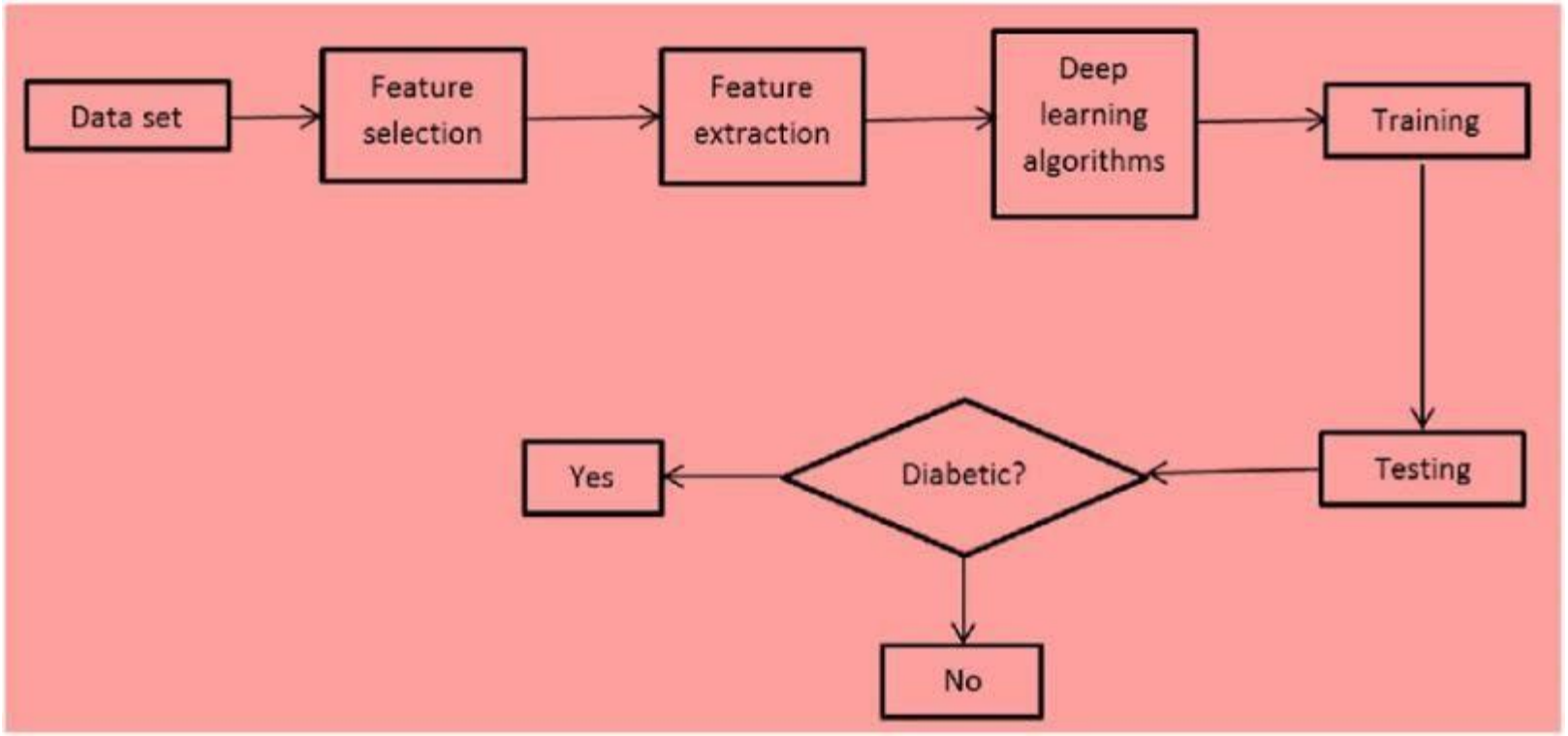
Provide high-level design view of the system.



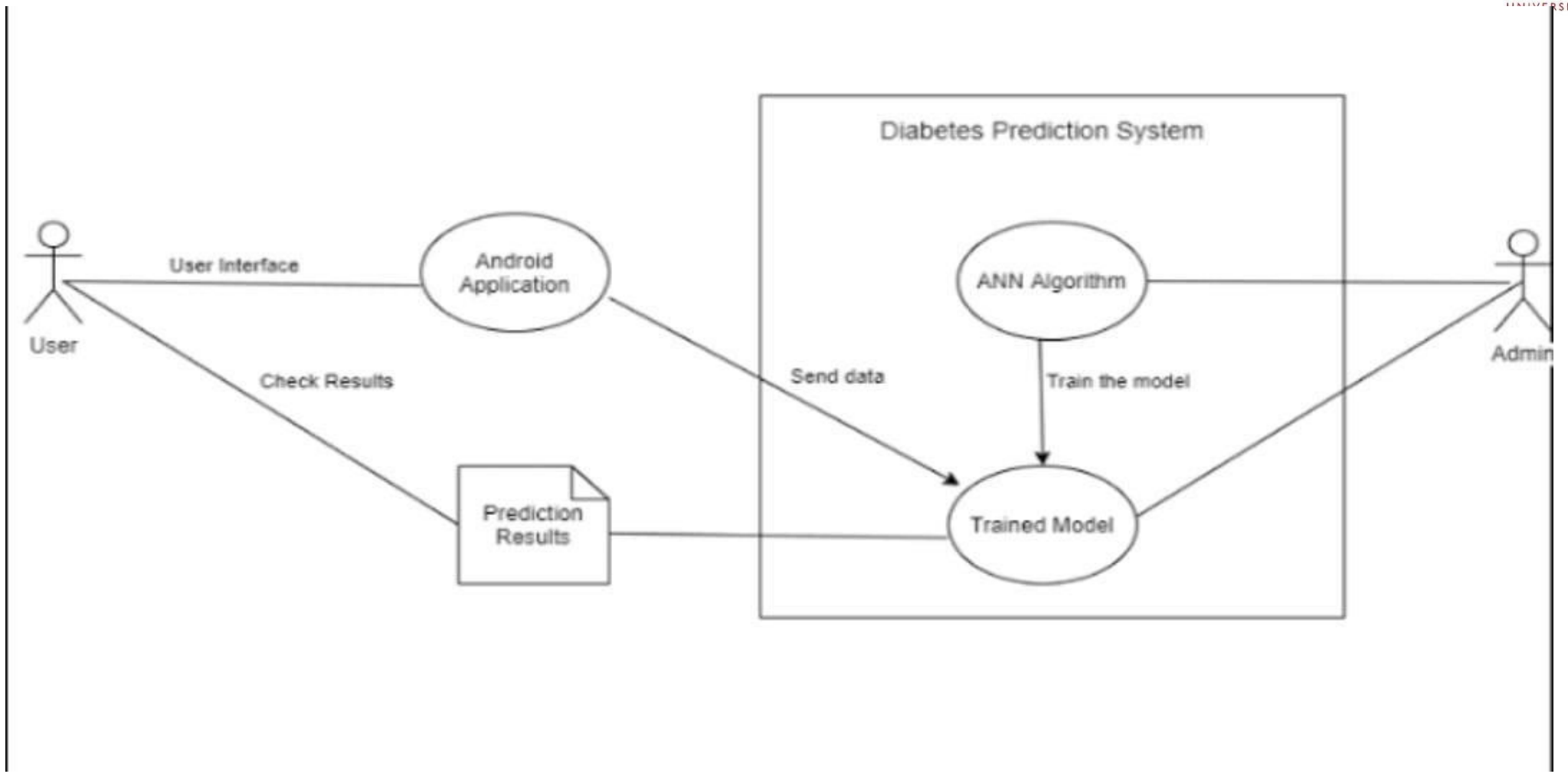
Master class diagram



ER Diagram

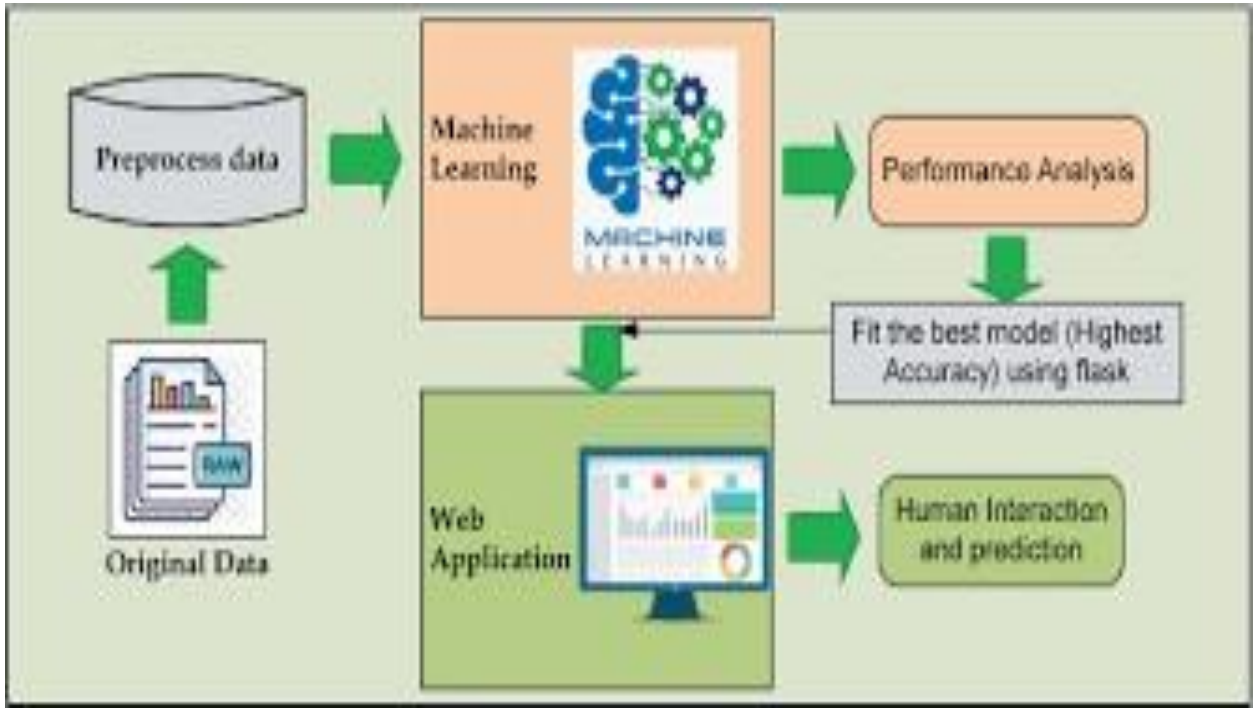

Data set
Feature selection
Feature extraction
Deep learning algorithms
Training
Testing
Diabetic?
Yes
No

User Interface Diagrams/ Use Case Diagrams



Report Layouts

External Interfaces



Technologies Used

What technologies you plan to use and why

The technologies used in diabetes prediction are:

- Support vector machine (SVM)
- Logistic Regression

All these algorithms together provide an accuracy of about 87% in diabetes prediction.

Project Progress

What is the project progress so far?

One fourth of the project has been completed so far with more towards the enchancing our knowledge about the topic and dependencies involved.

References

Provide references pertaining to your research according to IEEE format.

Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importances for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

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.

Nahla B., Andrew et al,"Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.

Thank You