Early Prediction of Diabetes

Abstract

Diabetes mellitus commonly known as diabetes is metabolic disease which causes high blood sugar. (Stephanie & Marina, 2020). Early detect of Diabetes and treatment helps to maintain good health of diabetic patients. Along with this, it also helps to reduce risk of other diseases like heart disease, brain stroke, limb amputation and mainly kidney failure. (DIABETES: A NATIONAL PLAN FOR ACTION. THE IMPORTANCE OF EARLY DIABETES DETECTION, 2004). In this project, objective is to detect diabetes at early stages by using common early-stage symptoms. Data preprocessed to remove any collinearity and found most contributing features or symptoms which leads to diabetes in future. Apriori algorithm used to find associated features and rules. Modelling methods such as logistic regression, K means clustering, Decision tree, Random Forest, Support Vector Clustering SVC used to predict early diabetes. Decision tree is utilized to visualize stepwise important features for efficient understand.

Introduction and Literature Review

Centers of Disease Control and prevention states that diabetic is long lasting chronic health condition which effects our body system generation of energy from food. (Diabetes, 2020) If person have diabetes, body does not generate enough insulin and more sugar stays in blood stream which leads to other diseases like kidney failure, heart disease and Vision loss. 34.2 million US adults have diabetes and 1 in 5 of them don't know that they have diabetes. Another factor is adult diabetes diagnosed has been doubled in last 20 years and it is 7th leading cause of death. (Diabetes, 2020)

A model of early prediction of diabetes 2019 paper (Alama, et al., 2019) worked on modelling by using diagnostic measurement such as Glucose , Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age and pregnancy factors. Dataset used was from National Institute of Diabetes and Digestive and Kidney Diseases. Paper shows there a strong association of diabetes with BMI and glucose level using Apriori method. Artificial neural network, random forest and K-means clustering techniques were implemented for the prediction of diabetes. The ANN technique provided a best accuracy of 75.7%. (Alama, et al., 2019)

Conference paper (Islam, Ferdousi, Rahman, & Bushra, 2019) worked on modelling using Symptoms which may cause diabetes. They used dataset that collected from the patients using direct questionnaire from Sylhet Diabetes Hospital of Sylhet, Bangladesh. They used Naive Bayes Algorithm, Logistic Regression Algorithm, and Random Forest Algorithm. After applying tenfold Cross- Validation and Percentage Split evaluation techniques, Random Forest has been found having best accuracy on this dataset. At the end, they created Web based tool for prediction of diabetes by taking symptoms as input. (Islam, Ferdousi, Rahman, & Bushra, 2019)

Executive Summary

Detection of diabetes at early stages provides critical health advantages to patients. In this project, Linear Vector Clustering gave highest sensitivity of 98.5% with accuracy 96.2% on testing dataset. Random Forest also provided highest sensitivity 98.5% but accuracy of 92.5%. Decision tree and Logistic regression also provided 97.6% and 98.4%Accuracy.

Decision tree modelling is utilized to visualize stepwise feature importance and observed polyuria is base root and with this base root Male gender had less probabilityof positive when compared to Female.

Associate rules and frequent features combinations detected using Apriori Algorithm,which provided Polyuria with Polydipsia and Polyuria with weakness Symptom combination have min support of 0.55 and min confidence of 0.8 in positive dataset.

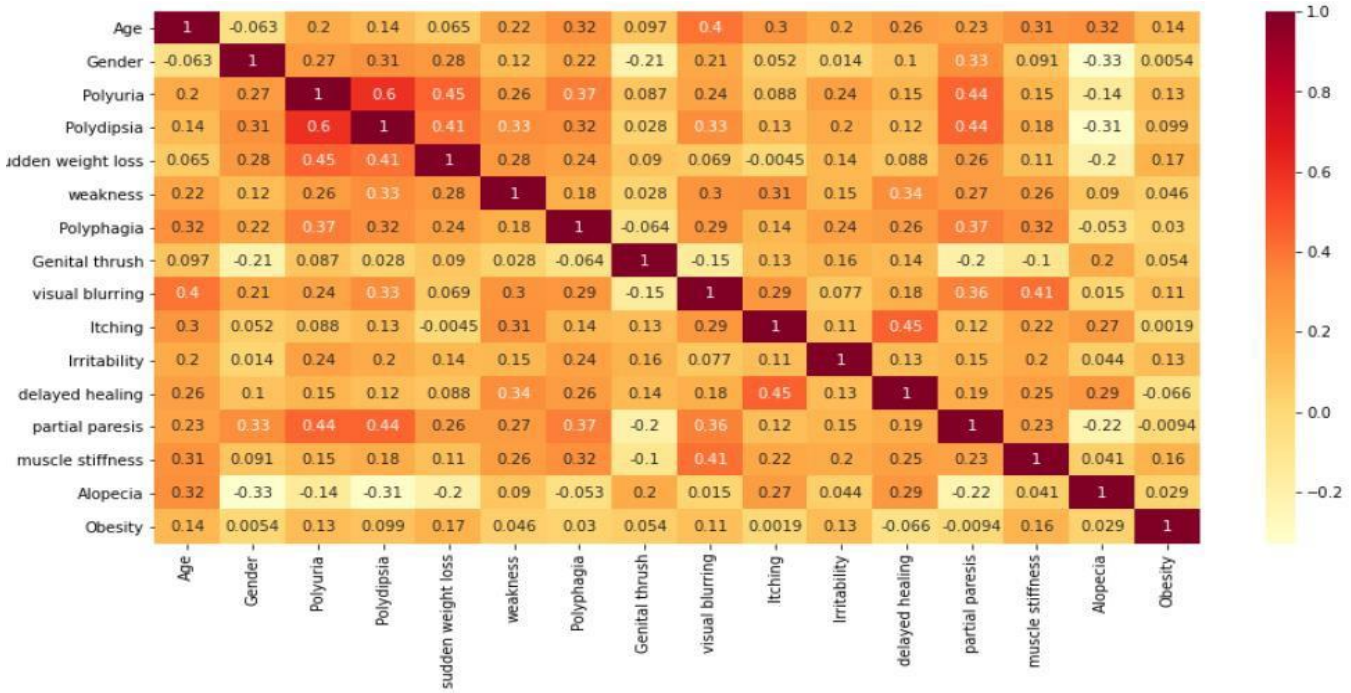
Data Preprocessing and Selected Methods

1. **Data Description**

   Early Diabetes detection dataset is collected using direct questionnaire to the patient from Sylhet Diabetes Hospital Sylhet, Bangladesh . Data set description states that questionnaire was asked to people who recently diagnosed with diabetes or who is nondiabetic but have symptoms. Dataset isfrom Center of Machine learning and intelligent Systems repository.

   There are 520 instances ,15 features and 1 output class-Positive or negative.

Figure 1 : Correlation Matrix of Features



Features	Feature Type	Description
Age	Continuous Variable	
Gender	Categorical , Male and Female	
Polyuria	Binary , Yes or No	Production of abnormally largevolumes of dilute urine
Polydipsia	Binary , Yes or No	Abnormally great thirst as asymptom of disease
Sudden weight loss	Binary , Yes or No	Sudden weight loss
Weakness	Binary , Yes or No	Weakness
Polyphagia	Binary , Yes or No	Medical term for excessive orextreme hunger
Genital Thrush	Binary , Yes or No	A yeast infection of the vagina andtissues at the opening of the vagina
Visual Blurring	Binary , Yes or No	Visual Blurring
Itching	Binary , Yes or No	Itching
Irritability	Binary , Yes or No	Irritability
Delayed healing	Binary , Yes or No	Delayed healing
Partial paresis	Binary , Yes or No	Weakening of a muscle or group ofmuscles
Muscle stiffness	Binary , Yes or No	Muscle stiffness
Alopecia	Binary , Yes or No	Sudden hair loss that starts with one or more circular bald patches that may overlap

Table 1 : Data features and Description

2. Data Preprocessing

In Data preprocessing step, data checked for any missing values and there were none. Then checked for any correlation between features. Polydipsia and polyuria resulted in correction value of 0.6 which is highest for entire data set. As 0.7 and above considered has high correlation and all features were considered for next step. Check for frequency distribution between two class to avoid any bias towards one class but there were 320 Positive and 200 Negative class.Features are standardized.

Model Selection Criteria

1. **Sensitivity** – Sensitivity is a measure of the proportion of actual positive cases that got predicted as positive. Sensitivity provides proportion of actual diabetic patients which are correctly classified, as this evaluate our objective which is to detect early diabetic more efficiently, it has been considered as primary metric for tunning.

$$Sentivity = \frac{TruePositive(Predicted\ Correct\ Positive)}{TruePositive + FalseNegative\ (Actual\ Positive)}$$

2. **Accuracy** – Accuracy indicates how overall model efficiency. It provides how many cases irrespectivepositive or negative class are detected correctly. This chosen to avoid bias towards positive class.

Models Selected and Tuning hyperparameter

Five classification models are used to predict diabetics whether positive or not. Data setdivided into Training, Tuning and Testing data set in ratio of 60:40:40 respectively. They are K Nearest neighbor with K-3, Linear SVM with C 0.7, Decision tree with depth 5, RandomForest with depth of 6 and Logistic regression with Inverse of regularization strength of 11.

Tuning dataset used for tuning hyperparameter for selected models. For hyperparameter selection, Sensitivity, Accuracy and F1 Score are used as evaluation metrics. Training data set used of modelling and testing data set used for final dataset to find best model.

Table 2 : Hyper Parameter Tuning Values

Method	Hyperparameter	Value
Logistic Regression	Inverse of regularization strength	7
Decision Tree	Max depth	5
Random Forest	Max depth	6
SVC	Penalty parameter	0.7
K Nearest Neighbor	Number of Neighbors K	3

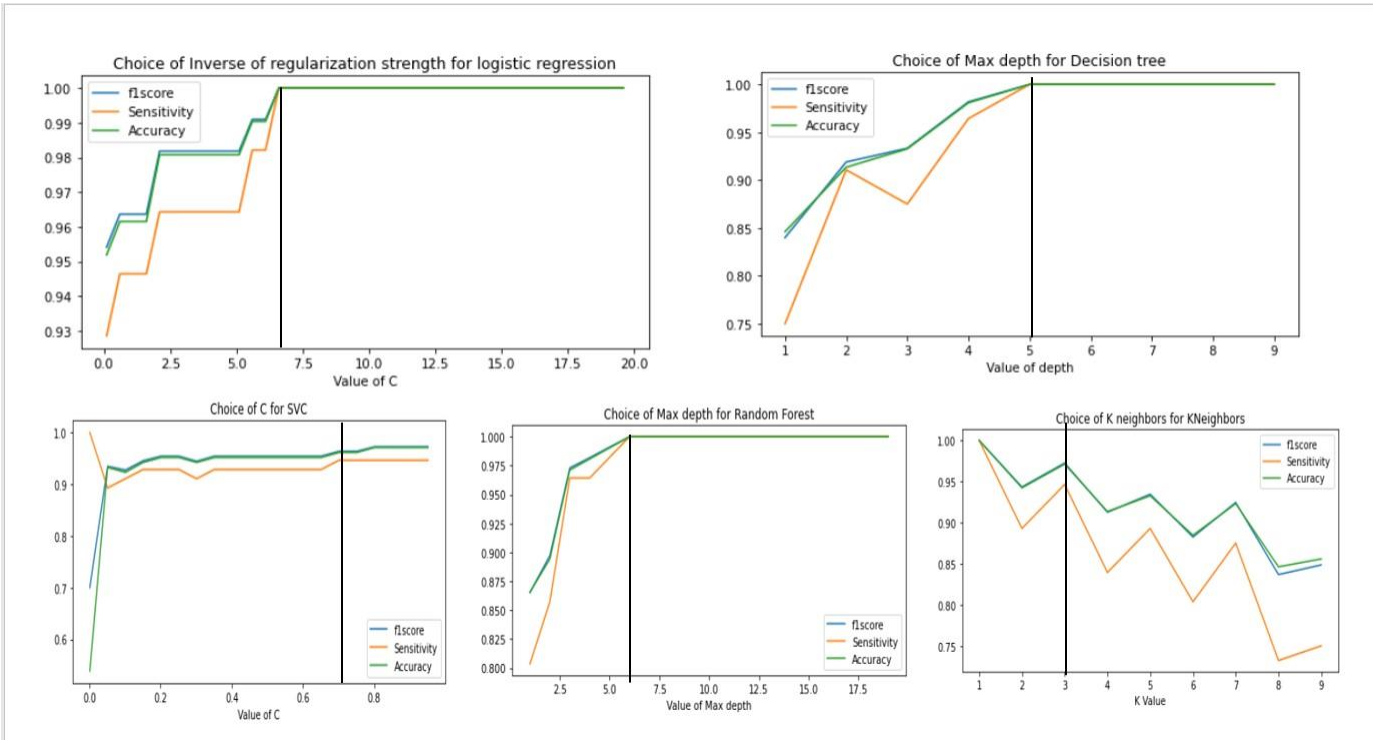


Figure 4 : Hyperparameter selection for all methods

Model Comparison (Training, Tuning, Validation) and Discussion

Testing dataset result shows that Linear SVM have high sensitivity of 98.5% with accuracy 97% and also training, tuning dataset got sensitivity of 96% and 91.1% withthis method. SVC generalized well with new dataset. Random forest has 98.5% Sensitivity with 92.3% accuracy and this method also generalized for new dataset.

Decision tree and Logistic regression have High Specificity, due to False positive accuracyand Sensitivity decreased.

Classifier	TP	FP	TN	FN	Precision	Recall	F1_score	Accuracy	Specificity	Sensitivity
Nearest Neighbors	61	1	38	4	0.984	0.938	0.961	0.952	0.974	0.938
Linear SVM	64	3	36	1	0.955	0.985	0.970	0.962	0.923	0.985
Decision Tree	62	0	39	3	1.000	0.954	0.976	0.971	1.000	0.954
Logistic Regression	63	0	39	2	1.000	0.969	0.984	0.981	1.000	0.969
Random Forest	64	7	32	1	0.901	0.985	0.941	0.923	0.821	0.985
Hybrid	64	6	33	1	0.914	0.985	0.948	0.933	0.846	0.985

Table 3: Result of Testing Dataset

Classifier	TP	FP	TN	FN	Precision	Recall	F1_score	Accuracy	Specificity	Sensitivity
Nearest Neighbors	193	1	112	6	0.995	0.970	0.982	0.978	0.991	0.970
Linear SVM	191	8	105	8	0.960	0.960	0.960	0.949	0.929	0.960
Decision Tree	192	2	111	7	0.990	0.965	0.977	0.971	0.982	0.965
Logistic Regression	186	0	113	13	1.000	0.935	0.966	0.958	1.000	0.935
Random Forest	189	12	101	10	0.940	0.950	0.945	0.929	0.894	0.950
Hybrid	189	12	101	10	0.940	0.950	0.945	0.929	0.894	0.950

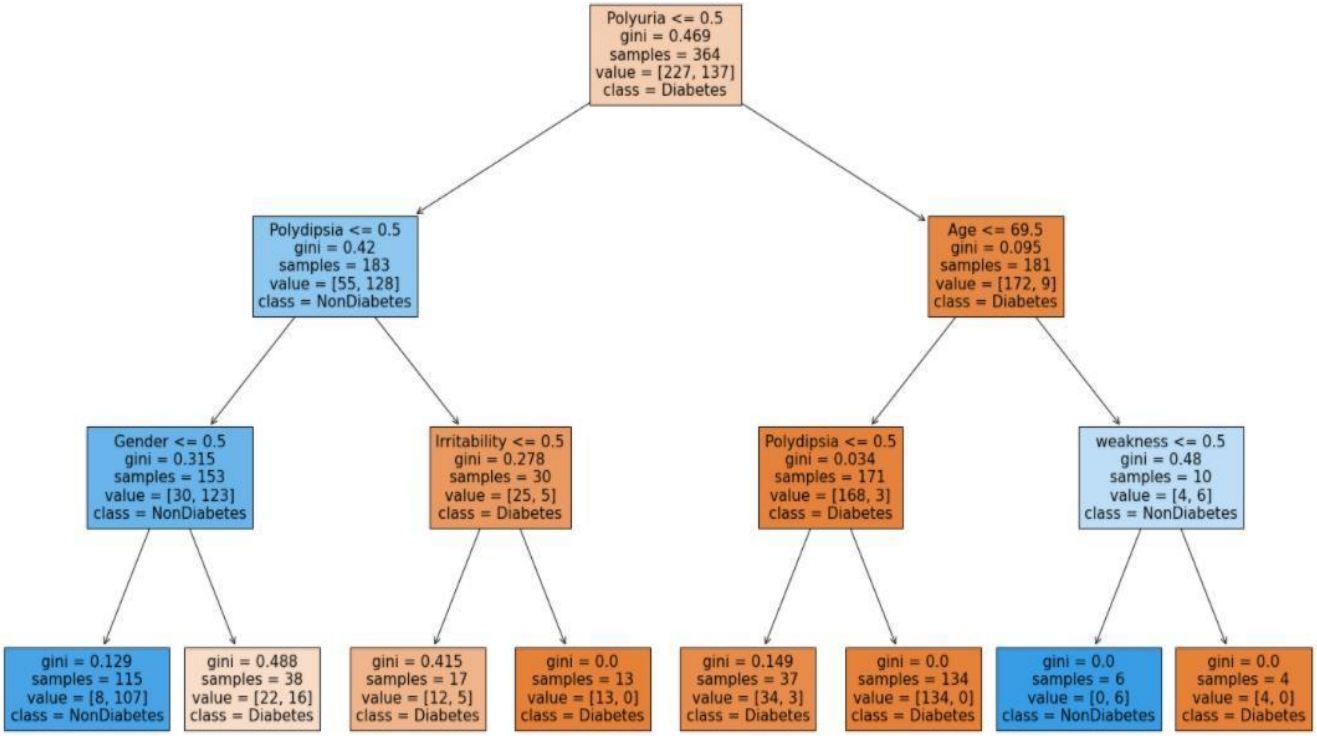
Table 4: Result of Training Dataset

Classifier	TP	FP	TN	FN	Precision	Recall	F1_score	Accuracy	Specificity	Sensitivity
Nearest Neighbors	54	0	48	2	1.000	0.964	0.982	0.981	1.000	0.964
Linear SVM	51	7	41	5	0.879	0.911	0.895	0.885	0.854	0.911
Decision Tree	50	1	47	6	0.980	0.893	0.935	0.933	0.979	0.893
Logistic Regression	52	1	47	4	0.981	0.929	0.954	0.952	0.979	0.929
Random Forest	49	5	43	7	0.907	0.875	0.891	0.885	0.896	0.875
Hybrid	49	5	43	7	0.907	0.875	0.891	0.885	0.896	0.875

Table 5: Result of Tuning Dataset

Decision tree Visualization for easy Understanding of Symptoms Relations

To visualize step wise features which are getting used to predict diabetes is gives understanding of important features. It will be helpful to decide about prediction class and it also indicates how many instances falls while deciding on any step. I used 3 step decision tree which provided 92.45% accuracy.



This 3-step decision tree gives brief answers for few questions.

1. If person does not have polyuria and polydipsia, does this guarantee Non diabetes? Left Chart shows on that next step is based gender, If Male then its Non diabetes but if Female then there is 11:8 Diabetes: Non-Diabetes chances. If Person have irritability, then there is high chance of Diabetes.

2. Another observation is even though person is polyuria and if age is >69.5 and there is no weakness then it shows its Nondiabetic. This shows old age people have polyuria which might be not strongly related to diabetes, but if person is <=69.5 and polyuria then it is early symptoms of diabetes.

Apriori Algorithm to find Frequent combination of Symptoms

Apriori Algorithm for finding frequent item sets and association rule learning. This Frequent item sets, and association rules helps to find patterns in symptoms for detection of positive class. For Apriori Algorithm, two parameters which need to decide are minimum support and minimum confidence. Minimum Support is what proportion of transaction should have feature or combination of features and it is 0.5 which is 55% transaction should have these feature. Minimum Confidence is how much percentage of confident data give when X feature is there in data will have Y feature also and it is provided as 90% confidence. Result shows that if person have weakness then there is 81.65% chance that Person will have Polyuria and in same way Polyuria given Polydipsia is 85.7% for Positive class.

```
20 freqItemSet, rules = apriori(list1, minSup=0.55, minConf=0.8)
21 freqItemSet
```

```
Out[90]: {1: {frozenset({'Polydipsia'}),
   frozenset({'weakness'}),
   frozenset({'Polyphagia'}),
   frozenset({'partial paresis'}),
   frozenset({'Polyuria'}),
   frozenset({'sudden weight loss'})},
  2: {frozenset({'Polyuria', 'weakness'}),
   frozenset({'Polydipsia', 'Polyuria'})}}
```

```
[91]: 1 rules
```

```
Out[91]: [[{'weakness'}, {'Polyuria'}, 0.8165137614678899],
 [{'Polydipsia'}, {'Polyuria'}, 0.8577777777777778]]
```

Conclusion and Future work

Early detection of diabetes helps people to avoid future critical health complication such as Kidney failure, heart disease. This project shows around 98% of actual Diabetes patients can be detected at early stages. Along with that Apriori showed Polydipsia, Polyuria, and weakness as frequent combination of Symptoms.

Limitation of this project is features are symptoms based on questionnaire. Along with this questionnaire, medical observation such as BMI, Glucose level in blood, Blood pressure etc. would have been produced more accuracy. Along with that number of instances are 520, as it is medical modelling larger datasets are always refereed for have good sensitivity result.

References

Alama, T. M., Iqbala, M. A., Alia, Y., Wahabb, A., Ijazb, S., Baig, T. I., . . . Abbas, Z. (2019). A model for earlyprediction of diabetes. *Informatics in Medicine Unlocked*.

Diabetes. (2020, July 11). Retrieved from Centers of Disease control and prevention:https://www.cdc.gov/diabetes/basics/diabetes.html

DIABETES: A NATIONAL PLAN FOR ACTION. THE IMPORTANCE OF EARLY DIABETES DETECTION. (2004, Dec 01). Retrieved from U.S. Department of Health & Human Services: https://aspe.hhs.gov/report/diabetes-national-plan-action/importance-early-diabetes-detection

Islam, M. M., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2019). Likelihood Prediction of Diabetes at Early StageUsing Data Mining Techniques. *Advances in Intelligent Systems and Computing book series.* AISC.

Stephanie, W., & Marina, B. (2020, Feb 26). *Everything You Need to Know About Diabetes*. Retrieved fromHealthline: https://www.healthline.com/health/diabetes

https://www.healthline.com/health/diabetes/facts-statistics-infographic#Types-of-diabetes https://www.cdc.gov/diabetes/data/statistics-report/diagnosed-undiagnosed-diabetes.html
https://www.sciencedirect.com/science/article/pii/S2352914819300176
https://www.phoebehealth.com/services/diabetes-center/diabetes-center-diabetes-symptoms
https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.