CSE-4212: Machine Learning and Data Mining Lab
Early Prediction of Diabetes Using Machine Learning Classifier

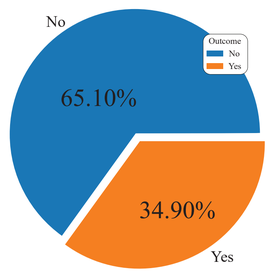
M.A Monaem Khan | Md. Abu Sama | Md. Alif Hossen Prince

Background

Diabetes, also known as chronic illness, is a group of metabolic diseases due to a high blood sugar level over a long period. The risk factor and severity of diabetes can be reduced significantly if a precise early prediction is possible. We are proposing a robust framework for diabetes prediction where filling the missing values, data standardization, and different Machine Learning(ML) classifiers (k-nearest Neighbour, Decision Trees, Random Forest, Naive Bayes, Support vector machine, Logistic R gression) are used. To improve the prediction of diabetes where the weights are estimated from the corresponding Area Under ROC Curve (AUC) of the ML model using the PIMA Indian Diabetes Dataset.

Dataset Description

Our PIMA Indian Diabetes Dataset[1] collected from Kaggle contains 768 rows and 9 columns. Here 500 rows are classified as 'No' and 268 as 'Yes.' Here, a pie chart is showing the ratio of yes no:



Overview Of our dataset[2] in table below:

SN	Columns Name	Description	Data Type	Null Row	Mean	Std	Median
1.	Pregnancies	Number of times pregnant	Int	0	3.84	3.37	3.0
2.	Glucose	Plasma glucose Concentration at 2 Hours in an Oral Glucose Tolerance Test	Int	0	120.89	31.97	117.0
3.	Blood Pressure	Diastolic Blood Pressure(mmHg)	Int	0	69.10	19.36	72.0
4.	Skin Thickness	Triceps Skin Fold Thickness(mm)	Int	0	20.54	15.95	23.0
5.	Insulin	2-Hour Serum Insulin(μU/ml)	Int	0	79.78	115.24	30.5
6.	BMI	Body Mass Index(Weight in kg/(Height in inches)2	Float	0	31.99	7.88	32.0
7.	Diabetes Pedigree Function	Diabetes Pedigree Function	Float	0	0.47	0.33	0.4
8.	Age	Age in years	Int	0	33.24	11.78	29.0

Methodology

Working Steps:

There are two phase for our work,

Phase-01: Training

Step 1: Loading Pima Indian Diabetes Dataset.
Step 2: Preprocessing the dataset.
Step 3: Dataset standardization.[2]
Step 4: Data splitting into 70 and 30 ratios.
Step 5: Train different ML classifiers by training the dataset.
Step 6: Evaluate the Trained ML classifier with the test dataset to find the best classifier for future prediction.

Phase-02: Prediction

Step 1: Load unlabeled data.
Step 2: Standardize unlabeled data.
Step 3: Load the best model.
Step 4: Make prediction.

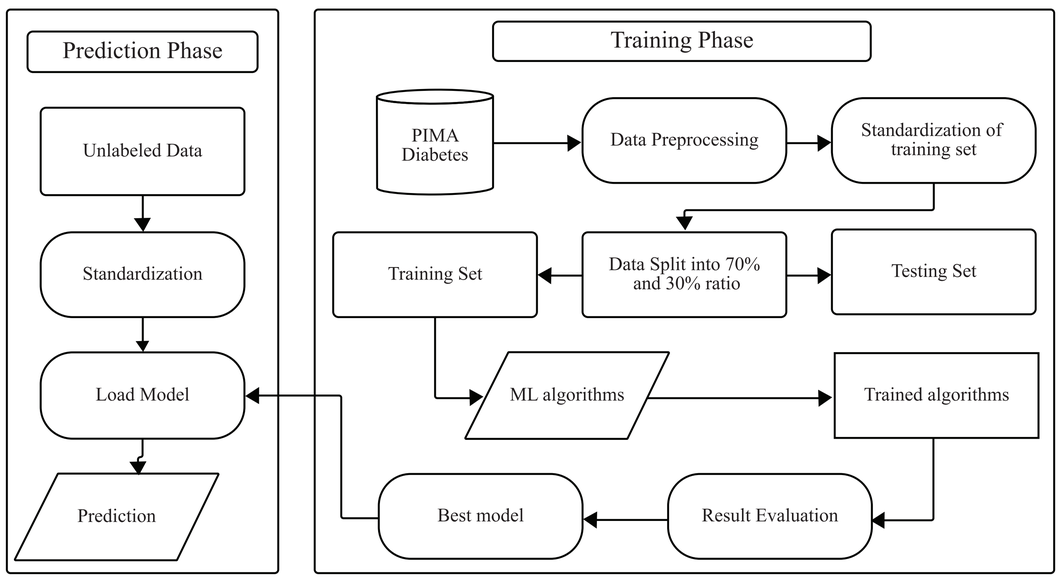
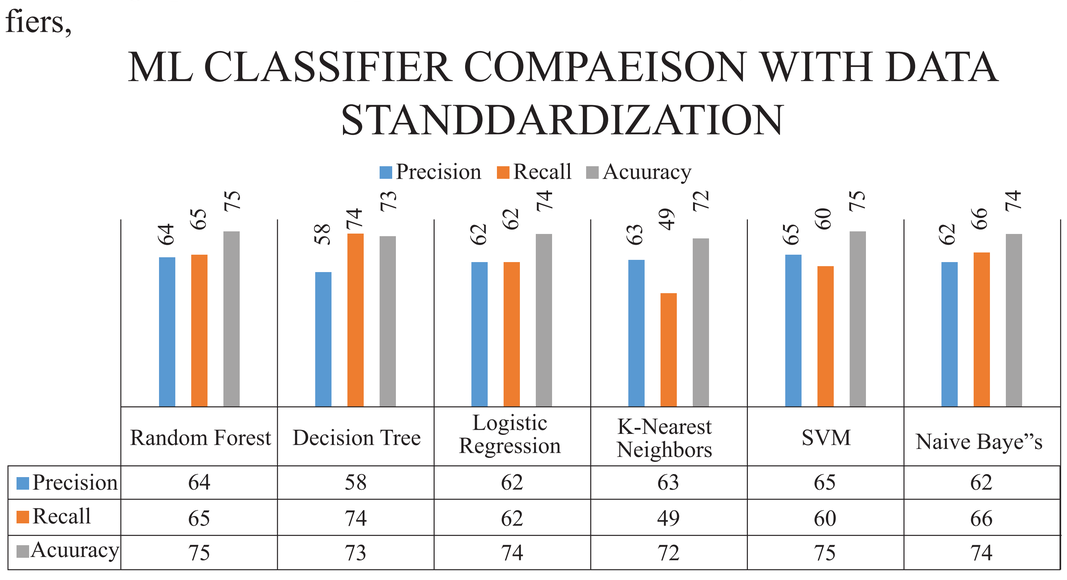


Fig: Diagram Of Our Working Procedure

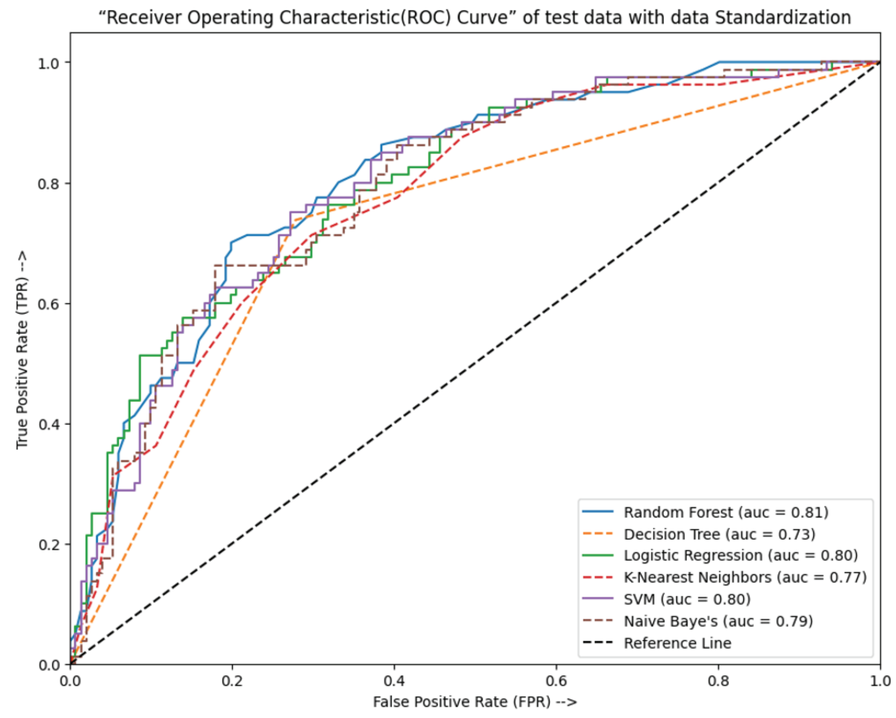
Experiments & Result

Here the graph shows the precision, recall and accuracy of different ML classifiers,



	Random Forest	Decision Tree	Logistic Regression	K-Nearest Neighbors	SVM	Naive Baye"s
Precision	64	58	62	63	65	62
Recall	65	74	62	49	60	66
Acuuracy	75	73	74	72	75	74

Here we show ROC-AUC of different ML Classifier,



According to both figures, we can see that Random Forest has the best accuracy, about 75% and the highest area under the curve (AUC), about 81%.

Here, we build a Graphical User Interface (GUI) by using our best ML classifier, so that anyone can check their diabetes risk.

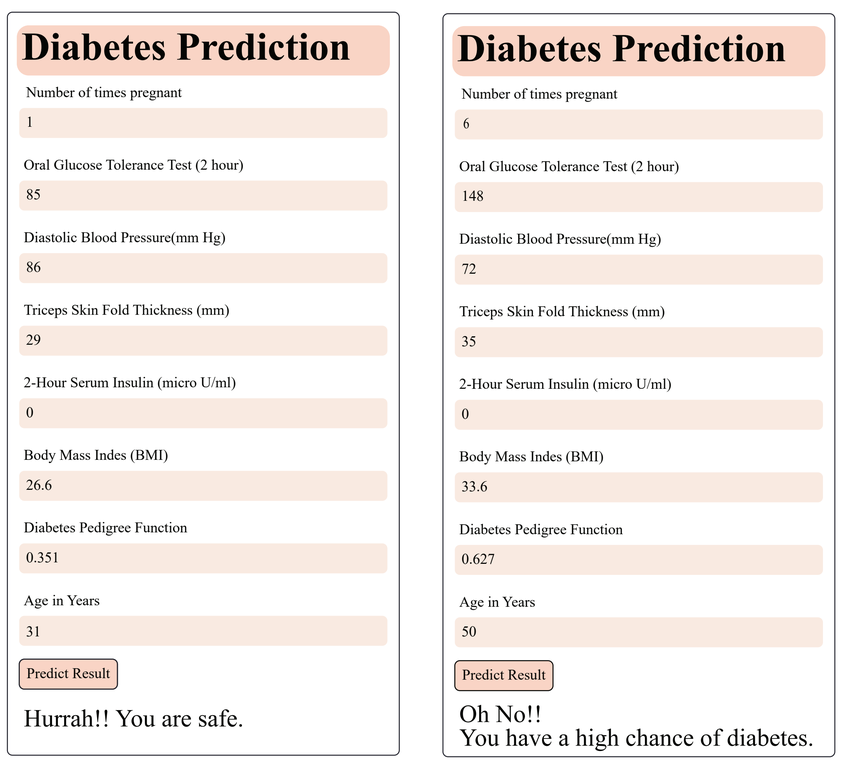


Fig: GUI for early Prediction of diabetes

Significance of the study

- It would help as a Decision Support System [DSS] for the hospital management, assisting them in making timely and quality decisions.
- It saves the hospital management the time and energy spent generating patient diseases in the existing system since most operations would automate under the proposed approach.

Future Work

In the future, the proposed trained model will be used to build a web app with a user-friendly interface. Additionally, the proposed framework will be applied to other medical contexts to verify their generality and versatility to predict the disease classes.

Discussion & Limitation

- In this work, diabetes prediction has been accomplished using the ML Classifier from the PID dataset. We select Random Forest as the best classifier based on accuracy and AUC for early prediction of diabetes.
- The main limitation of this study is that most the hospitals still keep medical records in manual form, making it almost impossible to get a local dataset to test the model that was formed.

Reference

[1] pima-indians-diabetes.csv. (n.d.). pima-indians-diabetes.csv | Kaggle. Retrieved January 9, 2023, from https:///datasets/kumargh/pimaindiansdiabetescsv.
[2] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516-76531.