PROJECT REPORT

Name of the Project: **PIMA Indian Diabetes Prediction**

Technology Used:

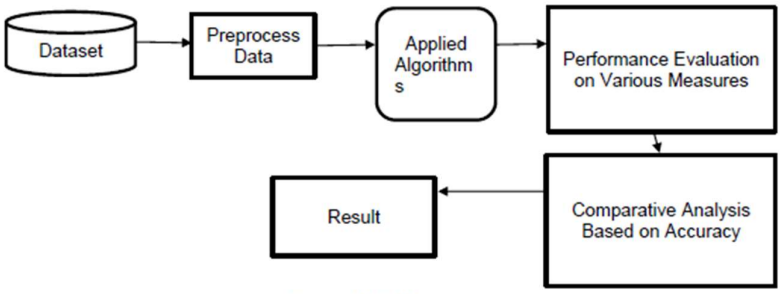
1. **Decision Tree Classifier:** Decision trees are versatile and intuitive machine learning models. They learn decision rules from the data and create a tree-like flowchart structure. Each internal node of the tree represents a decision based on a specific feature, and each leaf node represents a prediction or outcome. Decision trees are effective for both classification and regression tasks. They are interpretable, easy to understand, and can handle both numerical and categorical data.
2. **K-Nearest Neighbors (KNN):** KNN is a non-parametric, instance-based machine learning algorithm. It classifies new data points based on the majority class of their K nearest neighbors in the training set. KNN is simple to implement and works well with nonlinear decision boundaries. It can handle both classification and regression problems. KNN's main drawback is that it can be computationally expensive, especially with large datasets.
3. **Naive Bayes Classifier:** Naive Bayes is a probabilistic machine learning algorithm based on Bayes' theorem. It assumes that the features are conditionally independent given the class label, hence the "naive" assumption. Naive Bayes is fast, simple, and efficient for text classification and spam filtering tasks. It performs well with high-dimensional datasets and requires fewer training samples compared to other algorithms. However, Naive Bayes may oversimplify the relationships between features and class labels, sometimes leading to suboptimal performance.
4. **Random Forest Algorithm:** Random Forest is an ensemble learning algorithm that combines multiple decision trees. It creates an ensemble of decision trees, where each tree is trained on a random subset of the training data and a random subset of features. Random

Forest improves upon the individual decision trees' limitations by reducing overfitting and increasing predictive accuracy. It is robust against outliers and can handle high-dimensional data. Random Forest can be used for both classification and regression tasks and provides feature importance measures.

Project Description:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               2000 non-null   int64
 1   Glucose                   2000 non-null   int64
 2   BloodPressure             2000 non-null   int64
 3   SkinThickness             2000 non-null   int64
 4   Insulin                   2000 non-null   int64
 5   BMI                       2000 non-null   float64
 6   DiabetesPedigreeFunction  2000 non-null   float64
 7   Age                       2000 non-null   int64
 8   Outcome                   2000 non-null   int64
dtypes: float64(2), int64(7)
memory usage: 140.8 KB
```

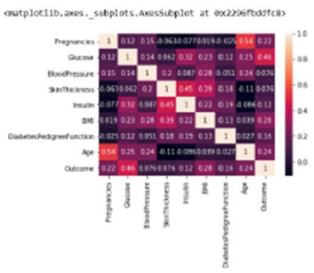
➔ There is no null values in dataset.



Proposed Model Diagram

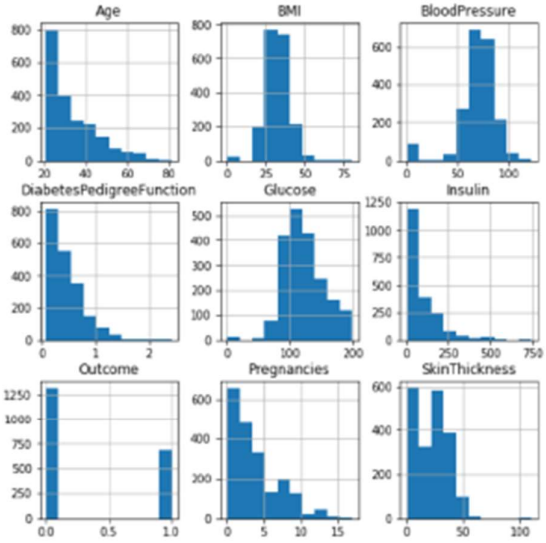
IV. RESULT & DISCUSSION

Correlation Matrix:



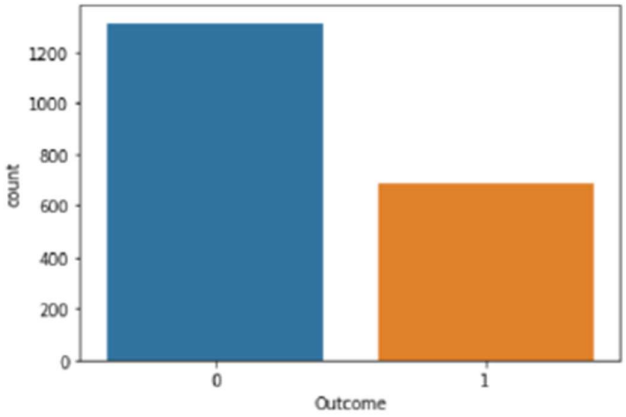
It is easy to see that there is no single feature that has a very high correlation with our outcome value. Some of the features have a negative correlation with the outcome value and some have positive.

Histogram:



Let's take a look at the plots. It shows how each feature and label is distributed along different ranges, which further confirms the need for scaling. Next, wherever you see discrete bars, it basically means that each of these is actually a categorical variable. We will need to handle these categorical variables before applying Machine Learning. Our outcome labels have two classes, 0 for no disease and 1 for disease.

Bar Plot For Outcome Class

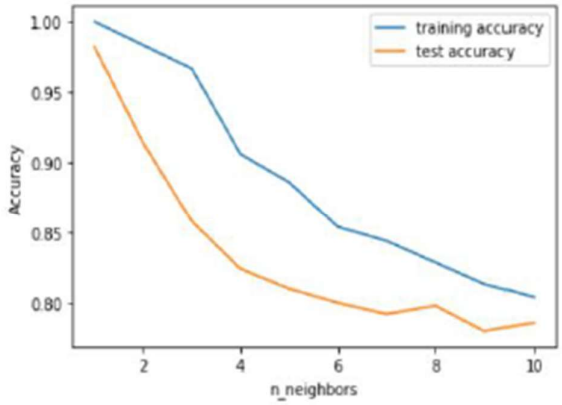


The above graph shows that the data is biased towards datapoints having outcome value as 0 where it means that diabetes was not present actually. The number of non-diabetics is almost twice the number of diabetic patients.

k-Nearest Neighbors:

The k-NN algorithm is arguably the simplest machine learning algorithm. Building the model consists only of storing the training data set. To make a prediction for a new data point, the algorithm finds the closest data points in the training data set, its "nearest neighbors."

First, let's investigate whether we can confirm the connection between model complexity and accuracy:



The above plot shows the training and test set accuracy on the y-axis against the setting of n_neighbors on the x-axis. Considering if we choose one single nearest neighbor, the prediction on the training set is perfect. But when more neighbors are considered, the training accuracy drops, indicating that using the single nearest neighbor leads to a model that is too complex. The best performance is somewhere around 9 neighbors.

Training Accuracy	0.81
Testing Accuracy	0.78

Table-1

Logistic regression:

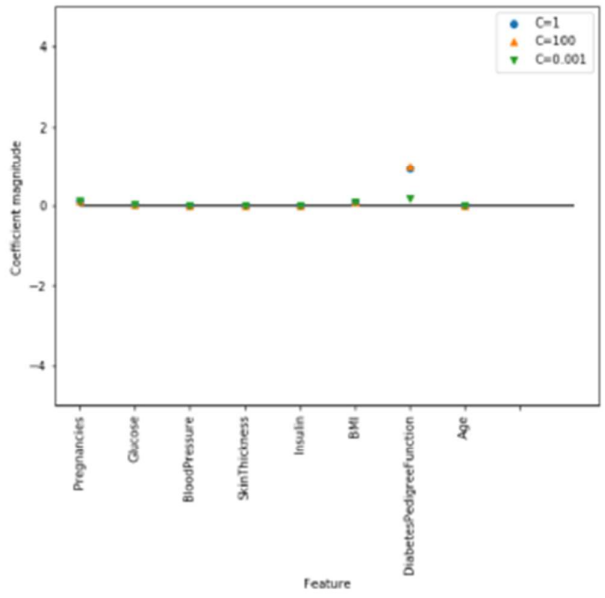
Logistic Regression is one of the most common classification algorithms.

	Training Accuracy	Testing Accuracy
C=1	0.779	0.788
C=0.01	0.784	0.780
C=100	0.778	0.792

Table-2

- In first row, the default value of C=1 provides with 77% accuracy on the training and 78% accuracy on the test set.
- In second row, using C=0.01 results are 78% accuracy on both the training and the test sets.
- Using C=100 results in a little bit lower accuracy on the training set and little bit highest accuracy on the test set, confirming that less regularization and a more complex model may not generalize better than default setting.

Therefore, we should choose default value C=1.



Decision Tree:

This classifier creates a decision tree based on which, it assigns the class values to each data point. Here, we can vary the maximum number of features to be considered while creating the model.

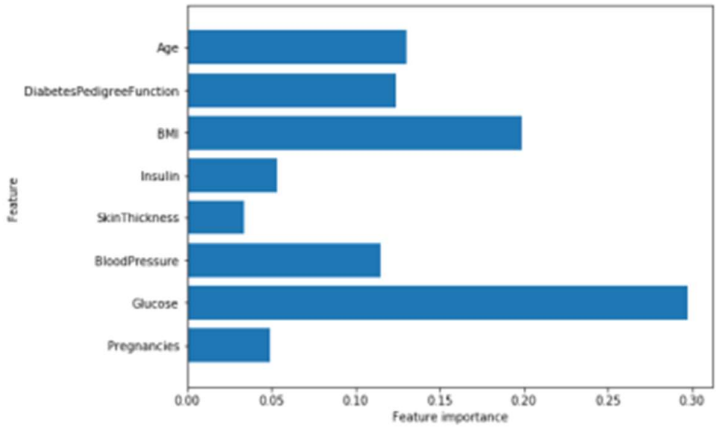
Training Accuracy	1.00
Testing Accuracy	0.99

Table-3

The accuracy on the training set is 100% and the test set accuracy is also good.

Feature Importance in Decision Trees

Feature importance rates how important each feature is for the decision a tree makes. It is a number between 0 and 1 for each feature, where 0 means "not used at all" and 1 means "perfectly predicts the target".



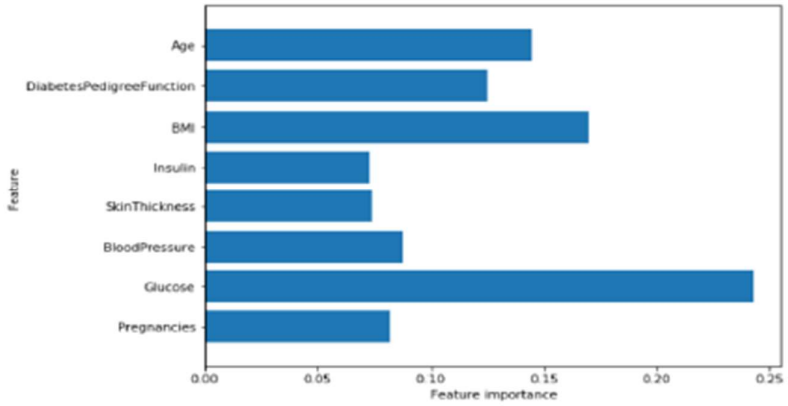
Feature "Glucose" is by far the most important feature.

Random Forest:

This classifier takes the concept of decision trees to the next level. It creates a forest of trees where each tree is formed by a random selection of features from the total features.

Training Accuracy	1.00
Testing Accuracy	0.974

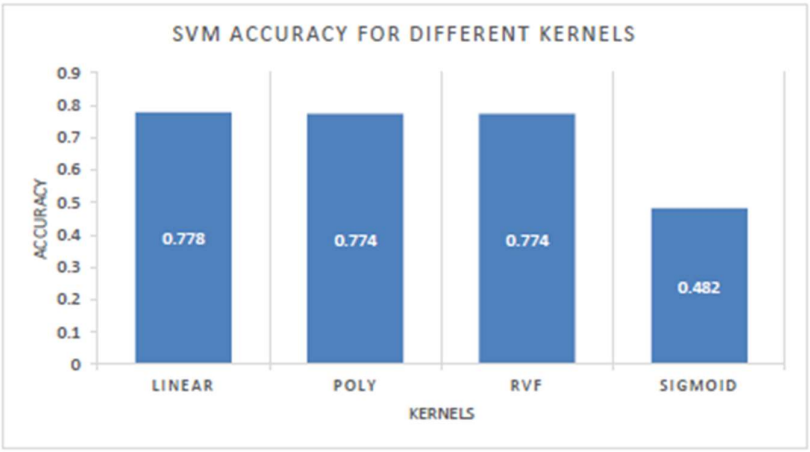
Feature importance in Random Forest:



Similarly to the single decision tree, the random forest also gives a lot of importance to the "Glucose" feature, but it also chooses "BMI" to be the 2nd most informative feature overall.

Support Vector Machine:

This classifier aims at forming a hyper plane that can separate the classes as much as possible by adjusting the distance between the data points and the hyper plane. There are several kernels based on which the hyper plane is decided. I tried four kernels namely, linear, poly, rbf, and sigmoid.



As can be seen from the plot above, the linear kernel performed the best for this dataset and achieved a score of 77%.

Accuracy Comparison:

Algorithms	Training Accuracy	Testing Accuracy
k-Nearest Neighbors	81%	78%
Logistic Regression	78%	78%
Decision Tree	98%	99%
Random Forest	94%	97%
SVM	76%	77%

Table-5

Table-5 shows the accuracy values for all five machine learning algorithms.

Table-5 shows that Decision Tree algorithm gives the best accuracy with 98% training accuracy and 99% testing accuracy.

Conclusion and Future Scope:

Diabetes is a serious and chronic condition. Diabetes can be detected early enough which can result in more effective treatment. This study also compares various classification models based on machine learning algorithms for predicting a patient's diabetic condition at the earliest feasible stage. After dataset balancing, classifiers' accuracy was compared. The prime objective of our research is to determine the early prediction of diabetes using the state of advanced MLA in one of the rural areas of North Kashmir. The data set employed for this experimentation was gathered from clinical professionals. In the medical diagnosis, we used a diabetes clinical data set with 403 instances and 11 attributes. The professionals (prediabetes specialists)

in the medical field have approved the features chosen for the early diagnosis of diabetes prediction. The prevalence of diabetes is showing an upward trend in Bandipora Kashmir. It is recommended that using state-of-the-art algorithms for early prediction can help in decreasing the upward trend of diabetes Six algorithms including RF, MLP, SVM, DT, GBC, and LR were utilized for this purpose all algorithms we achieved RF has the highest accuracy of 98%. RF also has produced successful outcomes for several statistical metrics including ROC Area, Recall, Precision, F-measure, and MCC. K-fold machine learning models such as cross-validation have been used to evaluate RF, MLP, SVM, DT, GBC, and LR. The framework utilized in this research will be applied to ensemble and hybridization Machine Learning to further recent research. In the future, a more comparison analysis between various datasets and their features can be conducted to identify all the crucial features for forecasting diabetes. To determine the best and most accurate diabetes prediction algorithm, a variety of various algorithms and combinations of algorithms can be examined.