Problem Statement

Diabetes is the second leading cause of death globally. After skin cancer, breast cancer is the most common cancer diagnosed in women over men. Diabetes occurs as a result of growth of sugar level commonly. Early Diagnosis of Diabetes can prove really helpful with respect to the number of increasing diabetic patients on average in the world.

Data Collection:

The Data Set has been collected from the open sourced data set of PIMA INDIAN PATIENTS available on Kaggle, (link: [https://www.kaggle.com/uc iml/pima-indians-diabetes-database](https://www.kaggle.com/uc%20iml/pima-indians-diabetes-database) ) diabetes.csv.

Data Overview:

This dataset consists several variables:

* **Pregnancies**: Number of times pregnant (Numerical)
* **Glucose**: Plasma glucose concentration 2 hours in an oral glucose tolerance test (Numerical)
* **BloodPressure**: Diastolic blood pressure (mm Hg) (Numerical)
* **SkinThickness**: Triceps skin fold thickness (mm) (Numerical)
* **Insulin**: 2-Hour serum insulin (mu U/ml) (Numerical)
* **BMI**: Body mass index (weight in kg/ (height in m) ^2) (Numerical)
* **DiabetesPedigreeFunction**: Diabetes pedigree function (Categorical)
* **Age**: Age (years) (Numerical)
* **Outcome**: Class variable (0 or 1) (Categorical)

Data Pre-processing:

Early improvement of the data set such as checking for null values, garbage values or empty values should be done so as to make Data Exploration as easy as possible.

In most of the real-world datasets, there are always a few null values. However, no model can handle these NULL or NaN values on its own. Therefore it is needed to intervene .On observation of the dataset for null values, it was detected some of the values present as (0’s) and thus they were replace as NaN values and then replace for the NaN value greater than 5% by their respective mean, modes.

Exploratory Data Analysis:

This section displays the summary statistic that quantitatively describes or summarizes features of a collection of information, the process of condensing key characteristics of the data set into simple numeric metrics. Some of the common metrics used are mean, standard deviation, and correlation.

From the visualizations as seen in the distribution plots, some features that have skewness,outliers and thus they were removed.

Also, a signification relationship was observed with respect to Outcome to be more likely a Diabetic for HIGH values of Glucose levels, Blood Pressure etc.

Model Building:

Before creating an model for the prediction all the features are standardized scale by removing the mean and scaling to unit variance by using sklearn’s StandardScaler() method. If our data has outliers, it can negatively impact our standardization, but have already removed them above and thus it helps our model.

KNN Algorithm:

Th K-Nearest Neighbors Classifier is one of the simplest yet most commonly used classifier in supervised machine learning, here we are going to use it as a comparative study of prediction with respect to decision trees algorithm. KNN is often considered as lazy classifier, it does not essentially trains the model to predict but it makes the classification of the class based on “Distance” between k sub-nearest neighbours found.

We have used ‘Elbow Method’ to identify the most optimal value of ‘K’ as seen in the code.

Highest accuracy achieved from K nearest neighbour Algorithm is 83%.

Decision Tree Algorithm:

Tree based learning algorithms are broad and popular family of related non parametric supervised method for classification. The basis of tree-based learners is the decision tree wherein a series of decision rules are chained that produce the greatest decrease in impurity at a node, as seen in the Tree Visualization in the notebook. One of the reasons for popularity of tree-based models is their interpretability. In scikit-kit learn, DecisionTreeClassifier operates like other methods; after the model is trained using fir, we can use it for model to predict on new class.

In order to identify the best performing hyperparameters of the DecisioTreeClassifier, we have used grid search method, the parameters best chosen were {'criterion': 'entropy', 'max\_depth': 6}

The best accuracy for the decision tree is achieved as ~78%.

Future Scope:

This Machine Learning application can be easily implemented in various scenarios as the performance seen from Knn Algorithm’s Recall being 88% correct in cases. We can add new features in as and when required to make more sophisticated models and help more patients to identify diabetes at early stages.