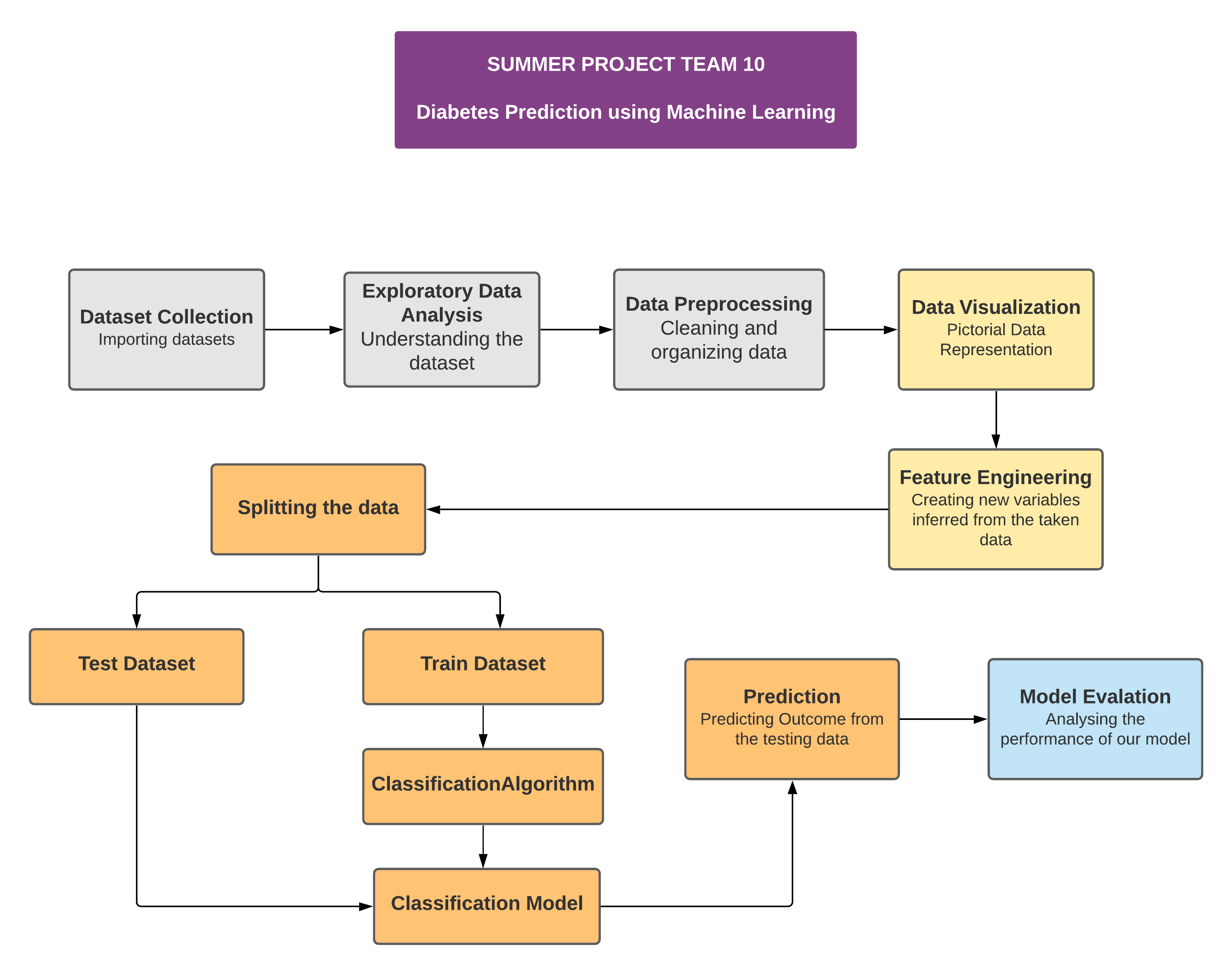
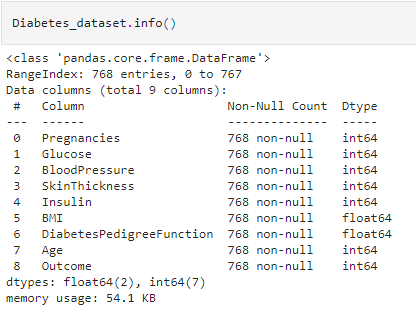
**ABSTRACT**

Diabetes is a chronic disease which increases blood sugar level and becomes severe when unnoticed or unidentified. The main aim of this paper is to build a better diabetes classification model (i.e.) to report whether the patient is diabetic or not. In this paper, we have proposed six different machine learning algorithm namely Logistic Regression, Support Vector Machine, Naïve bayes, Decision Tree, k-Nearest Neighbor and Random Forest to predict diabetes. We have performed data pre-processing to remove null values, outliers. Feature engineering is performed and new features are added to the dataset with the help of label encoding. Accuracy of all the six different algorithms is evaluated for both noisy as well as preprocessed data. Classification accuracy is improved with hyper parameter tuning using grid search method. In this proposed work we achieved highest accuracy of 100% on training data and 90.74% on testing data for Random Forest algorithm.

**PROPOSED WORK**

**DATASET COLLECTION & EXPLORATORY DATA ANALYSIS**

This research paper uses PIMA – Indians Diabetes dataset which is available in the UCI Machine Learning repository. The dataset contains 9 features and 768 records. Out of the 768 records 500 are Non diabetic and remaining 268 records are diabetic. EDA is the process of performing initial investigation on the dataset.

**DESCRIPTION OF ATTRIBUTES:**

**DATA PREPROCESSING**

In the first step data preprocessing is done. Preprocessing is the method by which we perform data cleaning i.e. raw dataset is converted into cleaned dataset. Missing values are replaced by the median values for improving the performance of the model. Outliers are removed with the help of Z score method. Features are scaled between 0 and 1 using Min-Max scaling technique.

**DATA VISULAIZATION**

Data visualization helps to understand the data and also explain the data to others. Histogram, density plot, box and whisker plot, correlation matrix plot, scatter matrix plot, pair plot have been plotted. In this paper we have found that there is a positive correlation between the features like: glucose level increases with age, insulin level increases with glucose level, skin thickness increases with insulin as well as with BMI. On plotting correlation matrix we found that glucose, BMI, Age is the most important features to identify whether the patient is diabetic or not.

**FEATURE ENGINEERING**

It is the process by which useful and relevant features are extracted from the raw dataset. It helps to avoid over fitting of the model. New features are created according to BMI, glucose and Insulin level with the help of label encoding method. Label encoding method converts the categorical variable into numeric form which is machine readable form.

**SPLITTING DATASET**

PIMA Indians Diabetes dataset is divided into 70% training data and 30% testing data.

**CLASSIFICATION ALGORITHM (MODEL BUILDING)**

This is one of the most important phases in the machine learning where we have implemented 6 algorithms. These algorithm include Logistic Regression, Support Vector Machine, Decision Tree, Naïve Bayes, K Nearest Neighbor, Random Forest.

**INPUT:** Data values

**OUTPUT:** Prediction is made. Accuracy, confusion matrix, classification report is displayed.

**PROCEDURE:**

ML\_algorithm\_used = [ LogisticRegression(), SVC(), DecisionTreeClassifier(), GuassianNB(), RandomForestClassifier(), KNeighborsClassifier() ]

Initialize variable ‘i’ to 1;

While (i!=7) do

Build the model using DiabetesModel = ML\_algorithm\_used[i]

Implement the DiabetesModel for training data using DiabetesModel.fit()

Implement DiabetesModel classify for testing data points using DiabetesModel.predict()

Print the confusion matrix and classification report

Calculate accuracy for test data, train data and display it.

end while

**HYPER-PARAMETER TUNING**

The parameters which is used define the model architecture are called as hyper parameters. Hyper parameter tuning is the process of finding the ideal model architecture. Accuracy of the build model is increased with hyper- parameter tuning. Grid search is the basic method used for hyper parameter tuning.

**EVALUATION**

Evaluation is done by measuring the classification accuracy of the built model using the formula:

ACCURACY = Total number of correct prediction made

------------------------------------------------------

Total number of predictions made