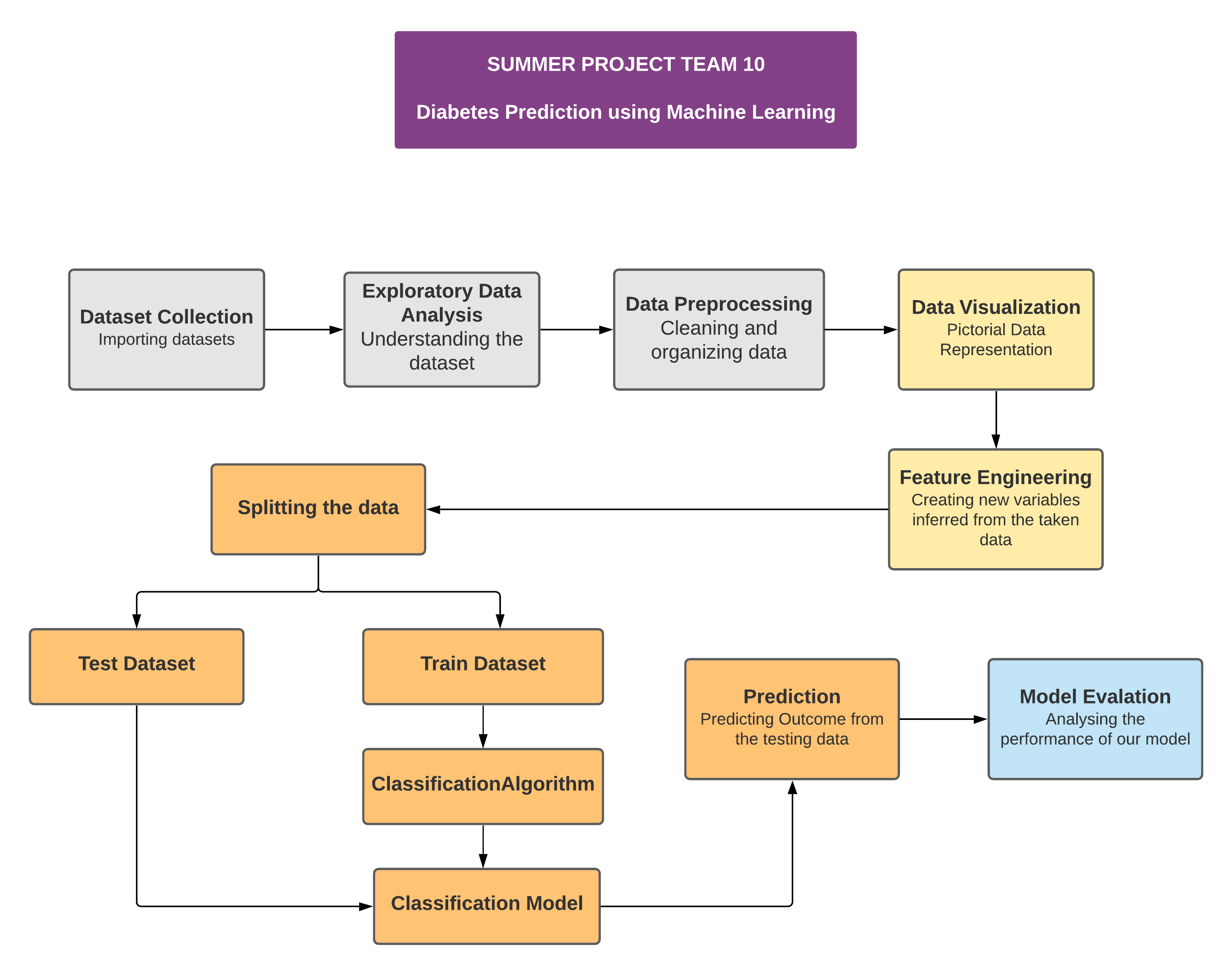
**DIABETES PREDICTION USING MACHINE LEARNING**

****

**ARCHITECTURE DIAGRAM**

**LINK:** [**click here**](https://lucid.app/lucidchart/bcea6828-af42-4db7-a298-568f02167e83/edit?viewport_loc=-34%2C-4463%2C2044%2C926%2C0_0&invitationId=inv_02d34c1f-353d-4b37-a4c9-2962eadde5b1)

Firstly we will import the dataset and do statistical analysis. Next we perform data visualization, feature engineering. We will split the dataset into Training data and testing data, 75% and 25% respectively. Here we use 5 different algorithm: Logistic Regression, Decision Tree, Support Vector Machine, Naïve Bayes and Random Forest and predict the result for the labeled test data. At last with the help of confusion matrix we evaluate the prediction made by the individual algorithms.

**HARDWARE REQUIREMENT**

A CPU minimum of 7th generation (Intel Core i7 processor) is recommended. A minimum of 4GB RAM machine is required for data not more than 10000 rows but 4GB and above is recommended higher dataset. SSD is also recommended for its speed and efficiency.

**SOFTWARE REQUIREMENT**

* **Google Colaboratory / Jupyter notebook**

Google Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education.

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.

* Python
* Anaconda software(optional)

**MACHINE LEARNING TOOLS**

* **numpy>=1.9.2**

NumPy arrays facilitate advanced mathematical and other types of operations on large numbers of data.

* **scipy>=0.15.1**

Scipy allows users to manipulate the data and visualize the data using a wide range of high-level Python commands.

* **scikit-learn>=0.18**

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

* **matplotlib>=1.4.3**

Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy.

* **pandas>=0.19**

Pandas is the most popular python library is used for data analysis. It provides highly optimized performance.

**DATASET IDENTIFIED:**

The dataset for Diabetes Prediction is collected from Kaggle. It comprises of 768 rows and 9 columns(features). Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on. Target variable is Outcome.

LINK: <https://www.kaggle.com/uciml/pima-indians-diabetes-database>

**TEAM MEMBERS & MODULE DIVISION:**

**RENITA V -** Data collection, Exploratory Data Analysis, Data Preprocessing

**SREVARDHANI R -** Data visualization, Feature Engineering

**SRUTHI S, PRUTHIEV A S –** Splitting data, Applying Algorithm- Classification model, Prediction

**LOGESHWARAN K -** Evaluation