****

# DEPARTMENT OF

**INFORMATION TECHNOLOGY**

A Project on

DIABETES PREDICTION USING MACHINE LEARNING

Batch – 12

|  |  |
| --- | --- |
| A.Siva Shankar | 21HP1A1255 |
| M.Sanjay Kumar | 22HP5A1206 |
| M.Naga Lokesh | 21HP1A1239 |
| B.Prasanth | 21HP1A1245 |

Under the guidance of **Mr V Vidya Sagar sir**

Date : 30-03-2024

Signature :

# ABSTRACT

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes is a disease caused due to the increase level of blood glucose. This high blood glucose produces the symptoms of frequent urination, increased thirst, and increased hunger. Diabetes is a one of the leading cause of blindness, kidney failure, amputations, heart failure and stroke. When we eat, our body turns food into sugars, or glucose. At that point, our pancreas is supposed to release insulin. Insulin serves as a key to open our cells, to allow the glucose to enter and allow us to use the glucose for energy. But with diabetes, this system does not work. Type 1 and type 2 diabetes are the most common forms of the disease, but there are also other kinds, such as gestational diabetes, which occurs during pregnancy, as well as other forms. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. The algorithms like K nearest neighbour, Logistic Regression, Random forest, Support vector machine and Decision tree are used. The accuracy of the model using each of the algorithms is calculated. Then the one with a good accuracy is taken as the model for predicting the diabetes

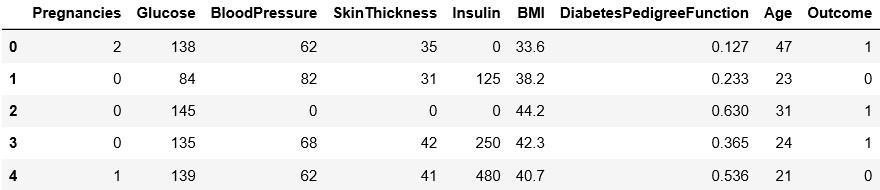
**Keywords:** Machine Learning, Diabetes, Decision tree, K nearest neighbour, Logistic Regression, Support vector Machine, Accuracy.

# INTRODUCTION

Diabetes is the fast growing disease among the people even among the youngsters. In understanding diabetes and how it develops, we need to understand what happens in the body without diabetes. Sugar (glucose) comes

from the foods that we eat, specifically carbohydrate foods. Carbohydrate foods provide our body with its main energy source everybody, even those people with diabetes, needs carbohydrate. Carbohydrate foods include bread, cereal, pasta, rice, fruit, dairy products and vegetables (especially starchy vegetables). When we eat these foods, the body breaks them down into glucose. The glucose moves around the body in the bloodstream. Some of the glucose is taken to our brain to help us think clearly and function. The remainder of the glucose is taken to the cells of our body for energy and also to our liver, where it is stored as energy that is used later by the body. In order for the body to use glucose for energy, insulin is required. Insulin is a hormone that is produced by the beta cells in the pancreas. Insulin works like a key to a door. Insulin attaches itself to doors on the cell, opening the door to allow glucose to move from the blood stream, through the door, and into the cell. If the pancreas is not able to produce enough insulin (insulin deficiency) or if the body cannot use the insulin it produces (insulin resistance), glucose builds up in the bloodstream (hyperglycaemia) and diabetes develops. Diabetes Mellitus means high levels of sugar (glucose) in the blood stream and in the urine

# METHODOLOGY

In this section we shall learn about the various classifiers used in machine learning to predict diabetes. We shall also explain our proposed methodology to improve the accuracy. Five different methods were used in this paper. The different methods used are defined below. The output is the accuracy metrics of the machine learning models. Then, the model can be used in prediction. Dataset Description The diabetes data set was originated from https://[www.kaggle.com/johndasilva/diabetes.](http://www.kaggle.com/johndasilva/diabetes) Diabetes dataset containing 2000 cases. The objective is to predict based on the measures to predict if the patient is diabetic or not.

\*The diabetes data set consists of 2000 data points, with 9 features each.

“Outcome” is the feature we are going to predict, 0 means No diabetes, 1 means diabetes.



Dataset

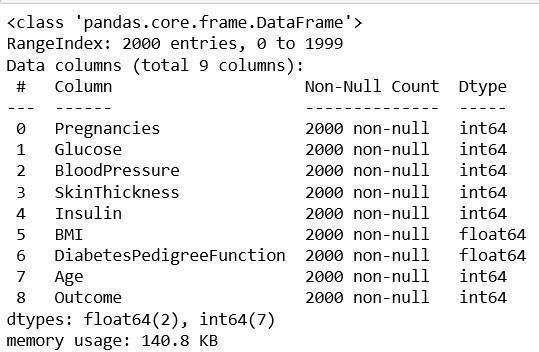
Preprocess Data

Applied Algorithm s

Performance Evaluation on Various Measures

Result

Comparative Analysis Based on Accuracy



# Literature Review

Person A uses the classification on diverse types of datasets that can be accomplished to decide if a person is diabetic or not. The diabetic patient’s data set is established by gathering data from hospital warehouse which contains two hundred instances with nine attributes. These instances of this dataset are referring to two groups

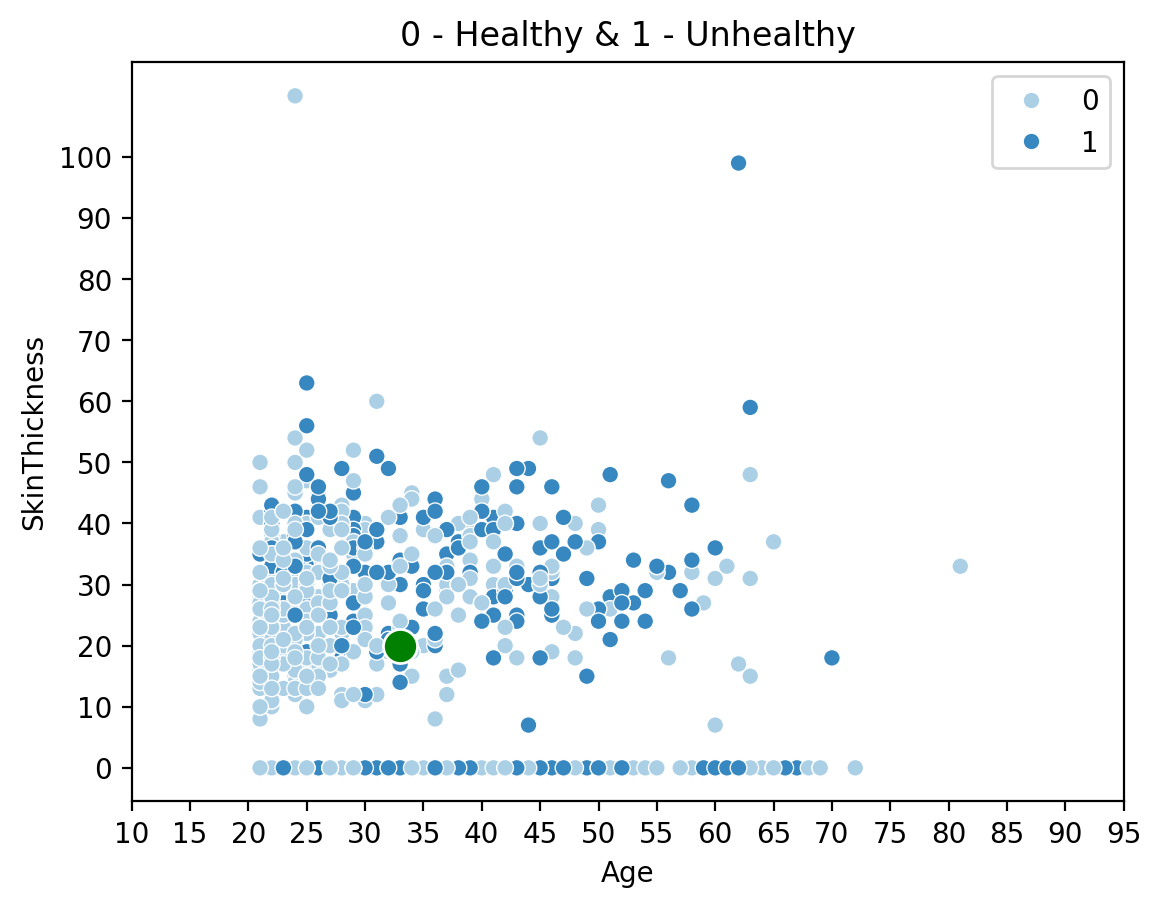
i.e. blood tests and urine tests. In this study the implementation can be done by using WEKA to classify the data and the data is assessed by means of 10-fold cross validation approach, as it performs very well on small datasets, and the outcomes are compared. The naïve Bayes, J48, REP Tree and Random Tree are used. It was concluded that J48 works best showing an accuracy of 60.2% among others.

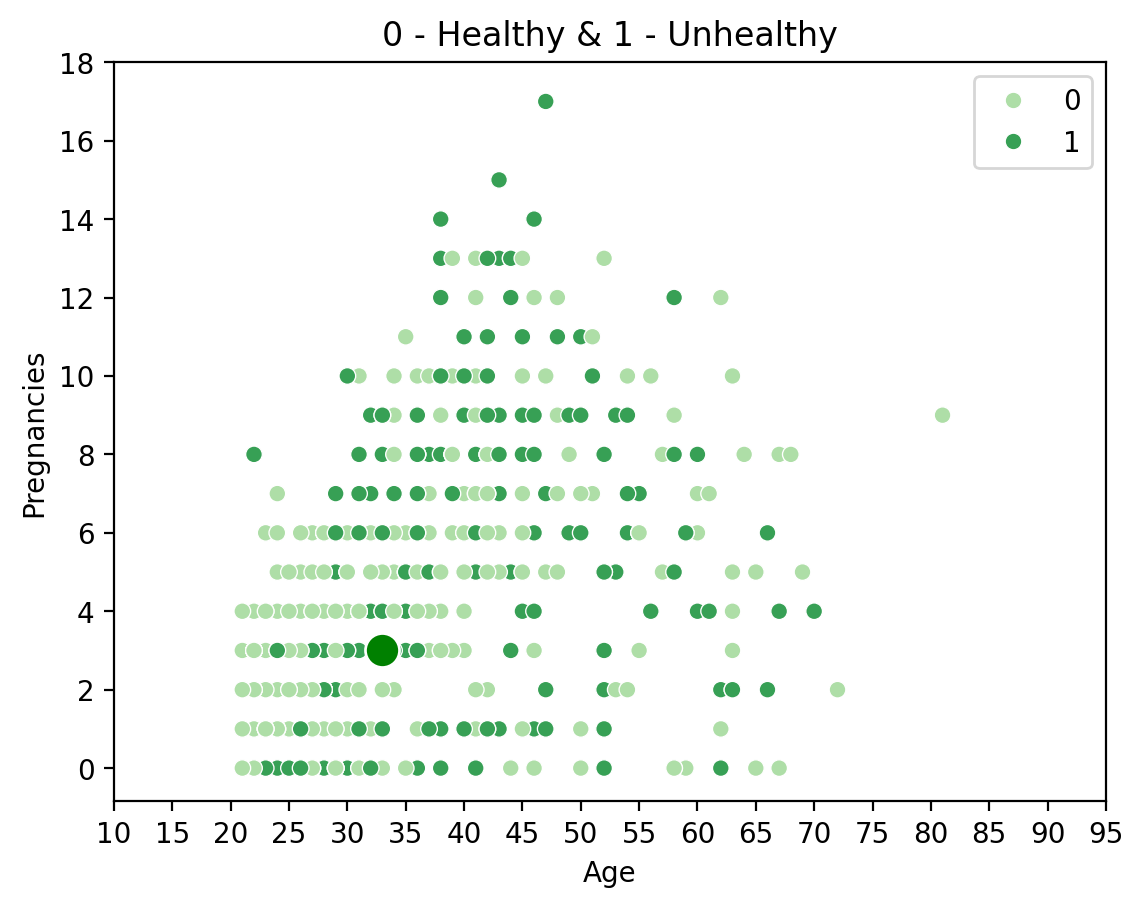
Person B aims to discover solutions to detect the diabetes by investigating and examining the patterns originate in the data via classification analysis by using Decision Tree and Naïve Bayes algorithms. The research hopes to propose a faster and more efficient method of identifying the disease that will help in well-timed cure of the patients. Using PIMA dataset and cross validation approach the study concluded that J48 algorithm gives an accuracy rate of 74.8% while the naïve Bayes gives an accuracy of 79.5% by using 70:30 split.

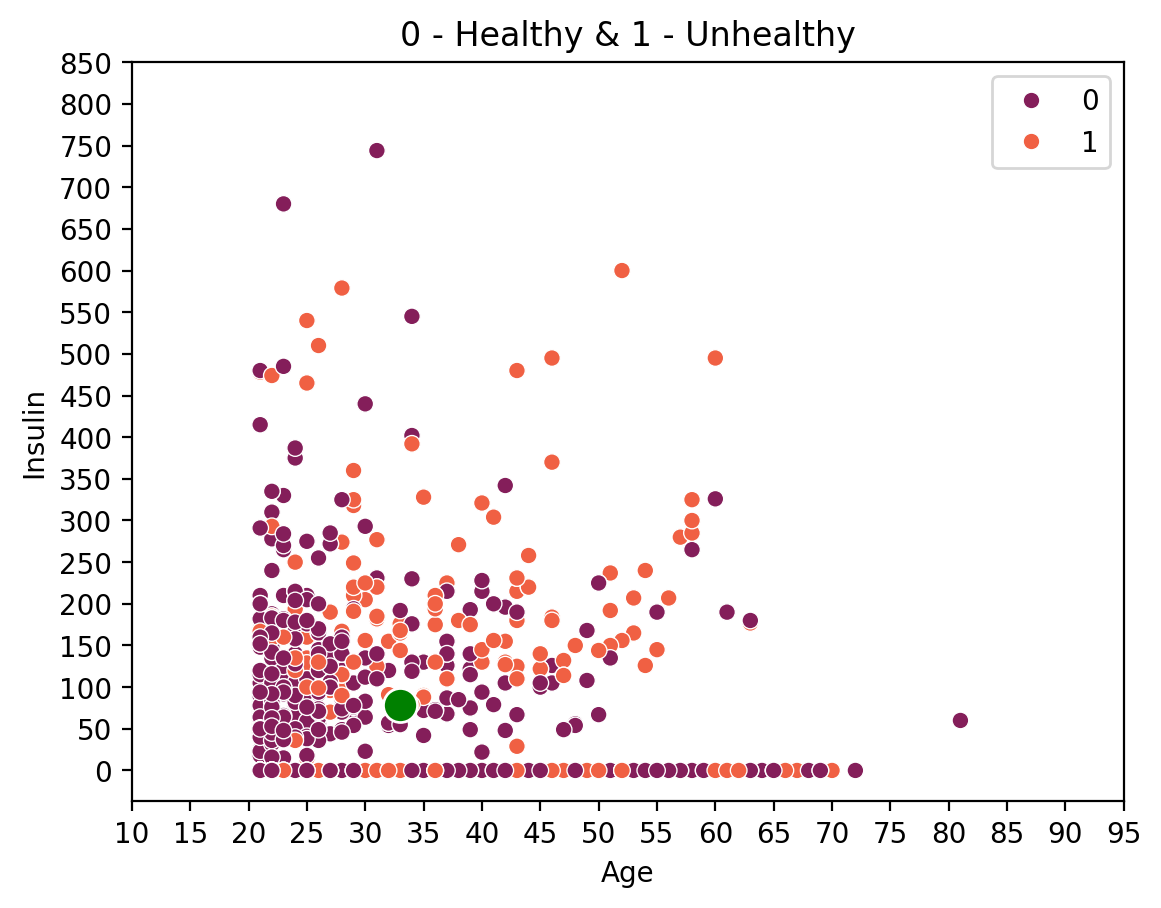
Person C aims to find and calculate the accuracy, sensitivity and specificity percentage of numerous classification methods and also tried to compare and analyse the results of several classification methods in WEKA, the study compares the performance of same classifiers when implemented on some other tools which includes Rapidminer and Matlabusing the same parameters (i.e. accuracy, sensitivity and specificity). They applied JRIP, Jgraft and BayesNet algorithms. The result shows that Jgraft shows highest accuracy i.e 81.3%, sensitivity is 59.7% and specificity is 81.4%. It was also concluded that WEKA works best than Matlab and Rapidminner.

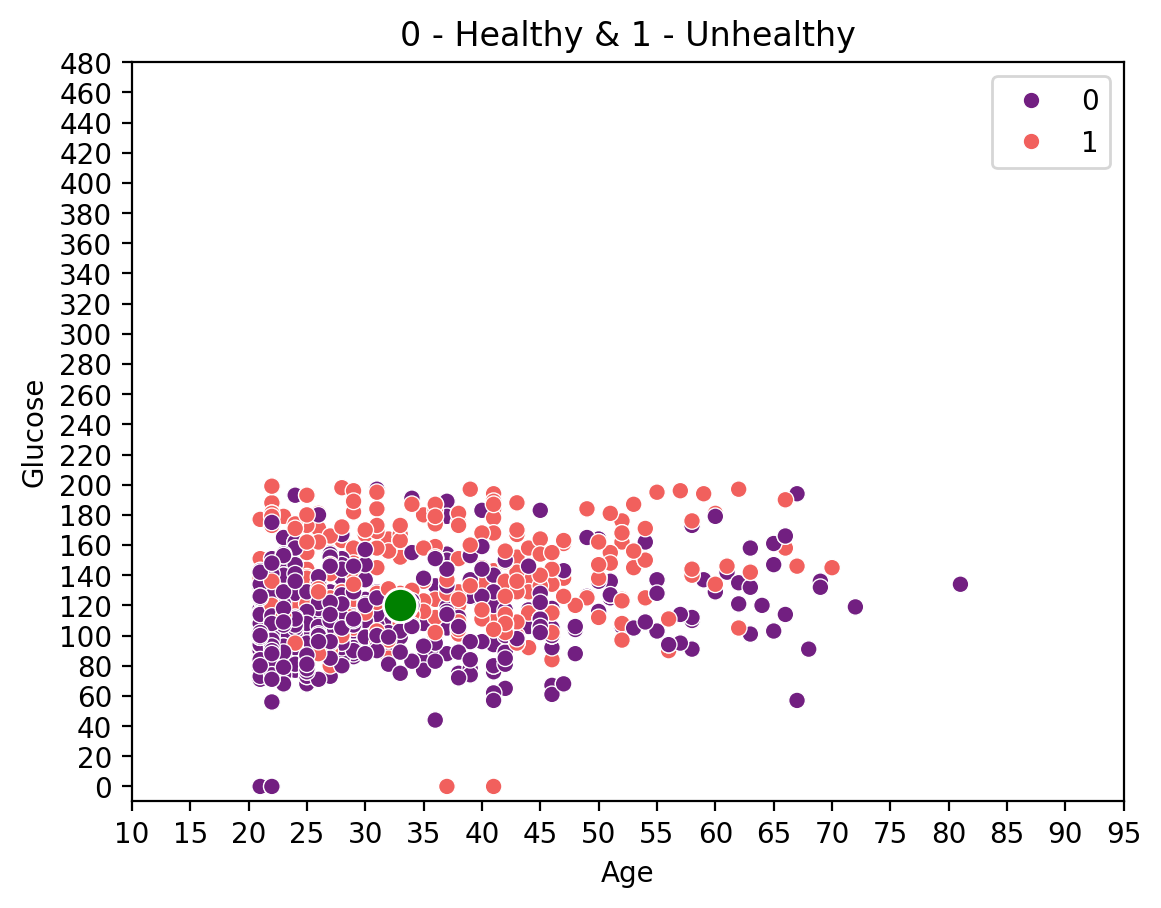
Person D focus on applying a decision tree algorithm named as CART on the diabetes dataset after applying the resample filter over the data. The author emphasis on the class imbalance problem and the need to handle this problem before applying any algorithm to achieve better accuracy rates. The class imbalance is a mostly occur in a dataset having dichotomous values, which means that the class variable have two possible outcomes and can be handled easily if observed earlier in data preprocessing stage and will help in boosting the accuracy of the predictive model.

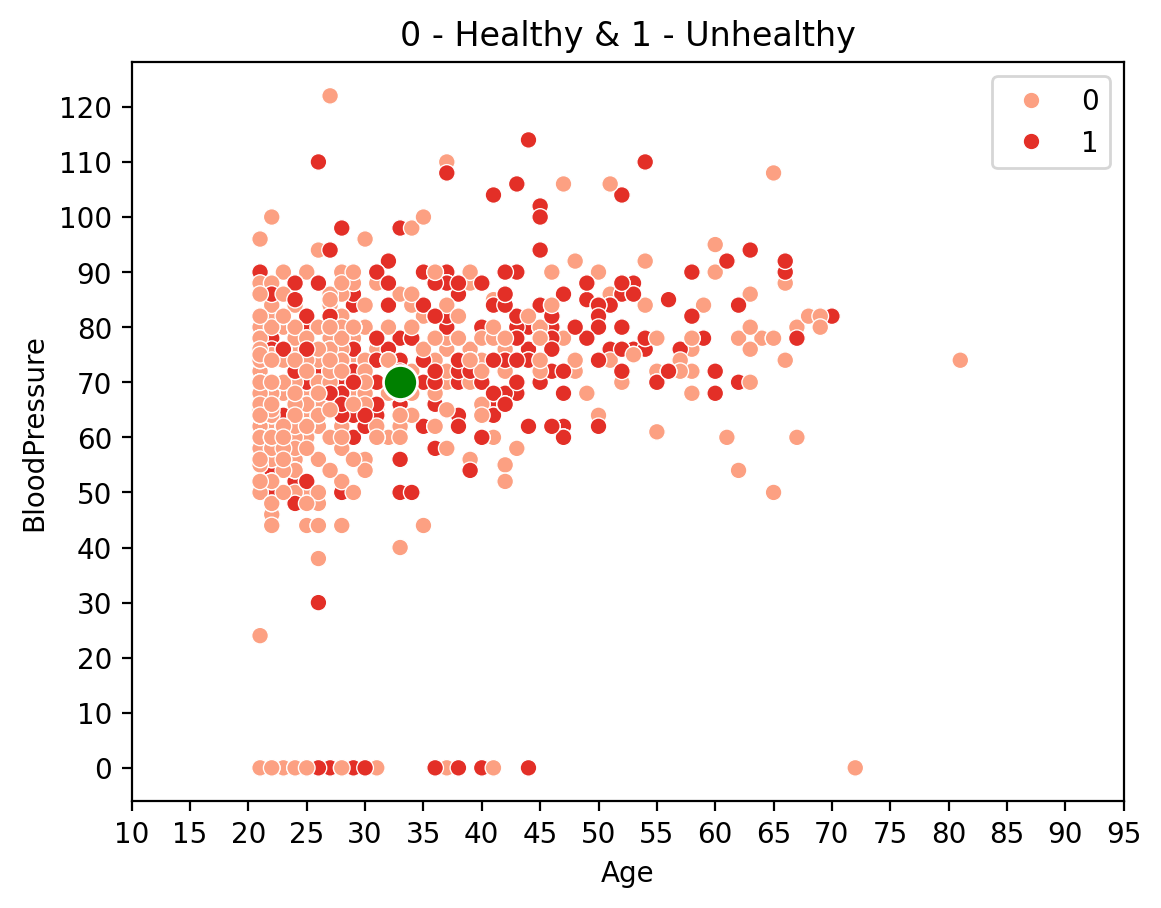
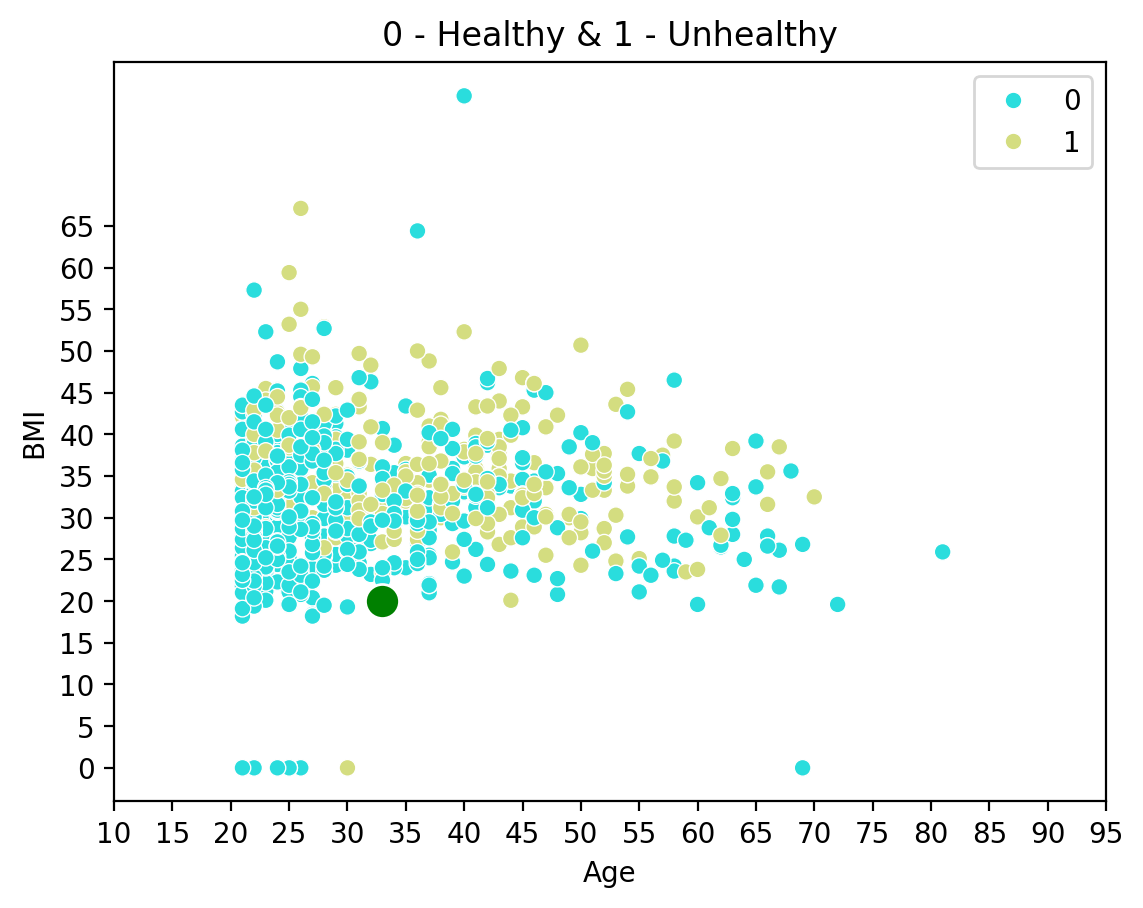
**Visualisation :**

****

****

****

****

****

# PROJECT CODE

# import streamlit as st

# import pandas as pd

# import numpy as np

# import matplotlib.pyplot as plt

# import seaborn as sns

# from sklearn.metrics import accuracy\_score

# from sklearn.ensemble import RandomForestClassifier

# from sklearn.model\_selection import train\_test\_split

# df = pd.read\_csv('diabetes.csv')

# # HEADINGS

# st.title('Diabetes Checkup')

# st.sidebar.header('Patient Data')

# st.subheader('Training Data Stats')

# st.write(df.describe())

# # X AND Y DATA

# x = df.drop(['Outcome'], axis=1)

# y = df['Outcome']

# x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=0.2, random\_state=0)

# # FUNCTION

# def user\_report():

# pregnancies = st.sidebar.text\_input('Pregnancies', 3)

# glucose = st.sidebar.text\_input('Glucose', 120)

# bp = st.sidebar.text\_input('Blood Pressure', 70)

# skinthickness = st.sidebar.text\_input('Skin Thickness', 20)

# insulin = st.sidebar.text\_input('Insulin', 79)

# bmi = st.sidebar.text\_input('BMI', 20)

# age = st.sidebar.text\_input('Age', 33)

# user\_report\_data = {

# 'Pregnancies': pregnancies,

# 'Glucose': glucose,

# 'BloodPressure': bp,

# 'SkinThickness': skinthickness,

# 'Insulin': insulin,

# 'BMI': bmi,

# 'Age': age

# }

# report\_data = pd.DataFrame(user\_report\_data, index=[0])

# return report\_data.astype(float)

# # PATIENT DATA

# user\_data = user\_report()

# st.subheader('Patient Data')

# st.write(user\_data)

# # MODEL

# rf = RandomForestClassifier()

# rf.fit(x\_train, y\_train)

# # Predict

# user\_result = rf.predict(user\_data)

# # VISUALISATIONS

# st.title('Visualised Patient Report')

# # COLOR FUNCTION

# if user\_result[0] == 0:

# color = 'green'

# else:

# color = 'red'

# # Age vs Pregnancies

# st.header('Pregnancy count Graph (Others vs Yours)')

# fig\_preg = plt.figure()

# ax1 = sns.scatterplot(x='Age', y='Pregnancies', data=df, hue='Outcome', palette='Greens')

# ax2 = sns.scatterplot(x=user\_data['Age'], y=user\_data['Pregnancies'], s=150, color=color)

# plt.xticks(np.arange(10, 100, 5))

# plt.yticks(np.arange(0, 20, 2))

# plt.title('0 - Healthy & 1 - Unhealthy')

# st.pyplot(fig\_preg)

# # Age vs Glucose

# st.header('Glucose Value Graph (Others vs Yours)')

# fig\_glucose = plt.figure()

# ax3 = sns.scatterplot(x='Age', y='Glucose', data=df, hue='Outcome', palette='magma')

# ax4 = sns.scatterplot(x=user\_data['Age'], y=user\_data['Glucose'], s=150, color=color)

# plt.xticks(np.arange(10, 100, 5))

# plt.yticks(np.arange(0, 500, 20))

# plt.title('0 - Healthy & 1 - Unhealthy')

# st.pyplot(fig\_glucose)

# # Age vs Bp

# st.header('Blood Pressure Value Graph (Others vs Yours)')

# fig\_bp = plt.figure()

# ax5 = sns.scatterplot(x='Age', y='BloodPressure', data=df, hue='Outcome', palette='Reds')

# ax6 = sns.scatterplot(x=user\_data['Age'], y=user\_data['BloodPressure'], s=150, color=color)

# plt.xticks(np.arange(10, 100, 5))

# plt.yticks(np.arange(0, 130, 10))

# plt.title('0 - Healthy & 1 - Unhealthy')

# st.pyplot(fig\_bp)

# # Age vs St

# st.header('Skin Thickness Value Graph (Others vs Yours)')

# fig\_st = plt.figure()

# ax7 = sns.scatterplot(x='Age', y='SkinThickness', data=df, hue='Outcome', palette='Blues')

# ax8 = sns.scatterplot(x=user\_data['Age'], y=user\_data['SkinThickness'], s=150, color=color)

# plt.xticks(np.arange(10, 100, 5))

# plt.yticks(np.arange(0, 110, 10))

# plt.title('0 - Healthy & 1 - Unhealthy')

# st.pyplot(fig\_st)

# # Age vs Insulin

# st.header('Insulin Value Graph (Others vs Yours)')

# fig\_i = plt.figure()

# ax9 = sns.scatterplot(x='Age', y='Insulin', data=df, hue='Outcome', palette='rocket')

# ax10 = sns.scatterplot(x=user\_data['Age'], y=user\_data['Insulin'], s=150, color=color)

# plt.xticks(np.arange(10, 100, 5))

# plt.yticks(np.arange(0, 900, 50))

# plt.title('0 - Healthy & 1 - Unhealthy')

# st.pyplot(fig\_i)

# # Age vs BMI

# st.header('BMI Value Graph (Others vs Yours)')

# fig\_bmi = plt.figure()

# ax11 = sns.scatterplot(x='Age', y='BMI', data=df, hue='Outcome', palette='rainbow')

# ax12 = sns.scatterplot(x=user\_data['Age'], y=user\_data['BMI'], s=150, color=color)

# plt.xticks(np.arange(10, 100, 5))

# plt.yticks(np.arange(0, 70, 5))

# plt.title('0 - Healthy & 1 - Unhealthy')

# st.pyplot(fig\_bmi)

# # OUTPUT

# st.subheader('Your Report: ')

# output = ''

# if user\_result[0] == 0:

# output = 'You are not Diabetic'

# else:

# output = 'You are Diabetic\nSuggestions to control Diabetes\n\nMedication Adherence \n\n\nPhysical Activity \n\n\nWeight Management \n\n\nStress Reduction \n\n\nSmoking Cessation \n\n\nAlcohol Moderation \n\n\nRegular Check-ups \n\n\nEducation and Empowerment'

# st.title(output)

# st.subheader('Accuracy: ')

# st.write(str(accuracy\_score(y\_test, rf.predict(x\_test)) \* 100) + '%')

# OUTPUT

# 

# 

# CONCLUSION AND FUTURE WORK

One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of diabetes. During this work, five machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on john Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 99% using Decision Tree algorithm. In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.