# REPORT ON

**Diabetes Healthcare: Comprehensive Dataset-AI using logistic regression** **project**

**Rupa Naga Lakshmi(2021BCSE07AED059)**

A mini project report submitted in partial fulfilment of the requirements for the degree

of

**BACHELOR OF TECHNOLOGY**

**Branch: COMPUTER SCIENCE AND ENGINEERING**

**Specialisation: AIML**

of Alliance University



**APRIL 2024**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**ALLIANCE COLLEGE OF ENGINEERING AND DESIGN**

ALLIANCE UNIVERSITY, BENGALURU

# ALLIANCE COLLEGE OF ENGINEERING AND DESIGN

(ALLIANCE UNIVERSITY, BENGALURU)

**Diabetes Healthcare: Comprehensive Dataset-AI using logistic regression** **project**

Bona fide record of work done by

**Rupa Naga Lakshmi (2021BCSE07AED047)**

A mini project report submitted in partial fulfilment of the requirements for the degree

of

**BACHELOR OF TECHNOLOGY**

## Branch: COMPUTER SCIENCE AND ENGINEERING

**Specialisation: AIML**

Of Alliance University

### April 2024

…………..……………….

### Dr. Chetan J Shelke

Faculty guide

Department of Computer Science and Engineering

Alliance College of Engineering and Design

**ABSTRACT:**

In this project, we want to use machine learning techniques to predict whether a patient has diabetes. This is important because early detection can help doctors treat diabetes and prevent complications. We use a dataset called the Pima Indians Diabetes Database (PIDD), which includes information about patients' health and whether they have diabetes.

To predict diabetes, we use different machine learning methods such as Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forests, and Gradient Boosting Classifiers. These methods help us analyze the data and identify patterns.

We measure how well each model works using metrics like Precision, Accuracy, Specificity, and Recall. By comparing the performance of these machine learning methods, we can find out which one works best for predicting diabetes.

Ultimately, the goal of this project is to help doctors and physicians detect diabetes early using machine learning techniques. This can improve patient care and lead to better health outcomes.

**INTRODUCTION:**

In the medical field, various classification strategies are employed to categorize data into different classes. Diabetes is a condition that disrupts the body's ability to produce the hormone insulin, leading to irregular carbohydrate metabolism and elevated blood glucose levels. High blood sugar is a common symptom of diabetes, which can result in severe complications if left untreated. Two significant complications include diabetic ketoacidosis and nonketotic hyperosmolar coma. Diabetes poses a serious health risk because of the body's inability to regulate blood sugar levels.

Diabetes is influenced by a range of factors such as height, weight, genetics, and insulin levels; however, the most crucial factor to consider is sugar concentration. Early identification of the issue is essential to avoid complications. The Pima Indians Diabetes Database (PIDD) from the National Institute of Diabetes and Digestive Diseases provides a dataset with various constraints.

The dataset is split into three sections for classification purposes: the training dataset, which is used to train the model; the validation dataset, which is utilized for fine-tuning parameters and assessing the model's performance in terms of accuracy and error rates between the training and validation datasets; and the testing dataset, which evaluates the model's output.

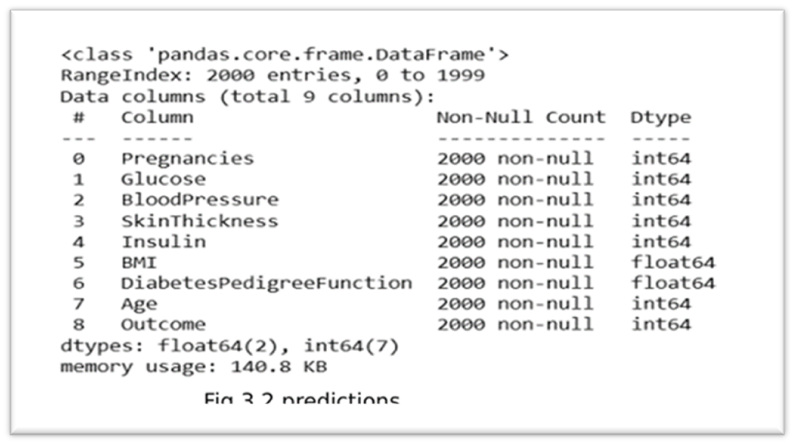
Several machine learning techniques are applied to the data, generating confusion matrices for comparison across different classification algorithms. This comparison helps determine which algorithm is most suitable for predicting diabetes. Correlation analyses between parameters and the highest accuracy score across various supervised machine learning algorithms are also conducted.

## The dataset collected is originally from the Pima Indians Diabetes Database and is available on Kaggle. It consists of several medical analyst variables and one target variable. The objective of the dataset is to predict whether the patient has diabetes or not. The dataset consists of several independent variables and one dependent variable, i.e., the outcome. Independent variables include the number of pregnancies the patient has had their BMI, insulin level, age, and so on as Shown in the Following table:

## 

➔ 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



➔ There are no null values in the dataset.

## METHODOLOGY:

**I] Dataset collection** – It includes data collection and understanding the data to study the

hidden patterns and trends which helps to predict and evaluate the results. Dataset carries

1405 rows i.e., the total number of data, and 10 columns i.e., the total number of features. Features include Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, DiabetesPedigreeFunction, Age

**II] Data Pre-processing:**

This phase of the model handles inconsistent data to get more

accurate and precise results like in this dataset ID is inconsistent so we dropped the feature.

This dataset doesn’t contain missing values. So, we imputed missing values for a few selected

attributes like Glucose level, Blood Pressure, Skin Thickness, BMI, and Age because these

attributes cannot have values of zero. The data was scaled using Standard Scaler. Since there

were a smaller number of features and important for prediction so no feature selection was

done.

**III]Missing value identification:**

Using the Panda library and SK-learn, we got the missing values in the datasets,

shown in Table 2. We replaced the missing value with the corresponding mean value. The link between numerical parameters ('age', 'BMI') and the target variable—which is distinguished by smoking status—is investigated using scatter plots.

**IV] Feature selection:**

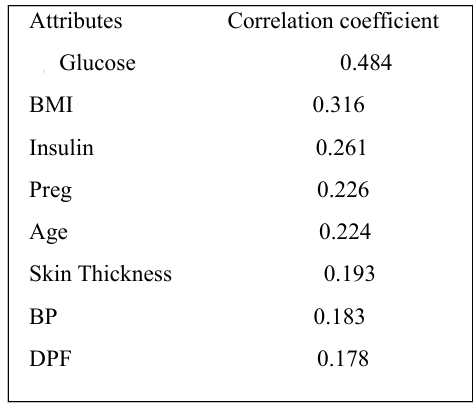
Pearson’s correlation method is a popular method to find the most

relevant attributes/features. The correlation coefficient is calculated in this method, which

correlates with the output and input attributes. The coefficient value remains in the range

between −1 and 1. The value above 0.5 and below −0.5 indicates a notable correlation, and

the zero value means no correlation

Fig 4.2 Correlation Table003A

**V] Scaling and Normalization:**

We performed feature scaling by normalizing the data from the 0 to 1 range, which boosted the

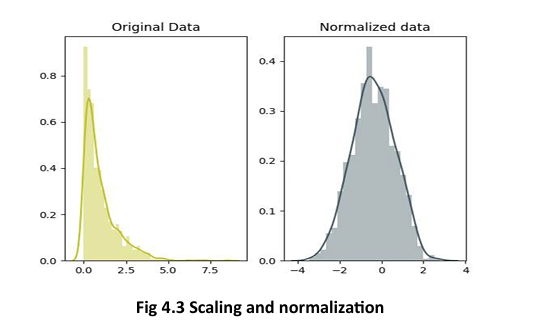
algorithm’s calculation speed.

scaling means that you're transforming your data so that it fits within a specific scale, like 0

100 or 0-1. You want to scale data when you're using methods based on measures of how far

apart data points are, like support vector machines (SVM) or k-nearest neighbours (KNN).

With these algorithms, a change of "1" in any numeric feature is given the same importance.

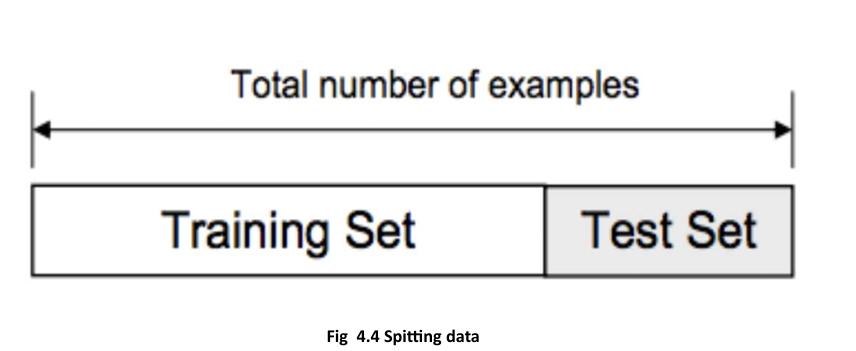


**VI] Splitting of data:**

After data cleaning and pre-processing, the dataset becomes ready to

train and test. In the train/split method, we split the dataset randomly into the training and

testing set. For Training, we took 1600 samples and for testing, we took 400 sample



**VII] Design and implementation of the classification model:**

In this research work, comprehensive studies are done by applying different ML

classification techniques like DT, RF, LR, and SVM.

**VIII] Machine learning classifier:**

We have developed a model using the Machine learning Technique. Used different classifiers and

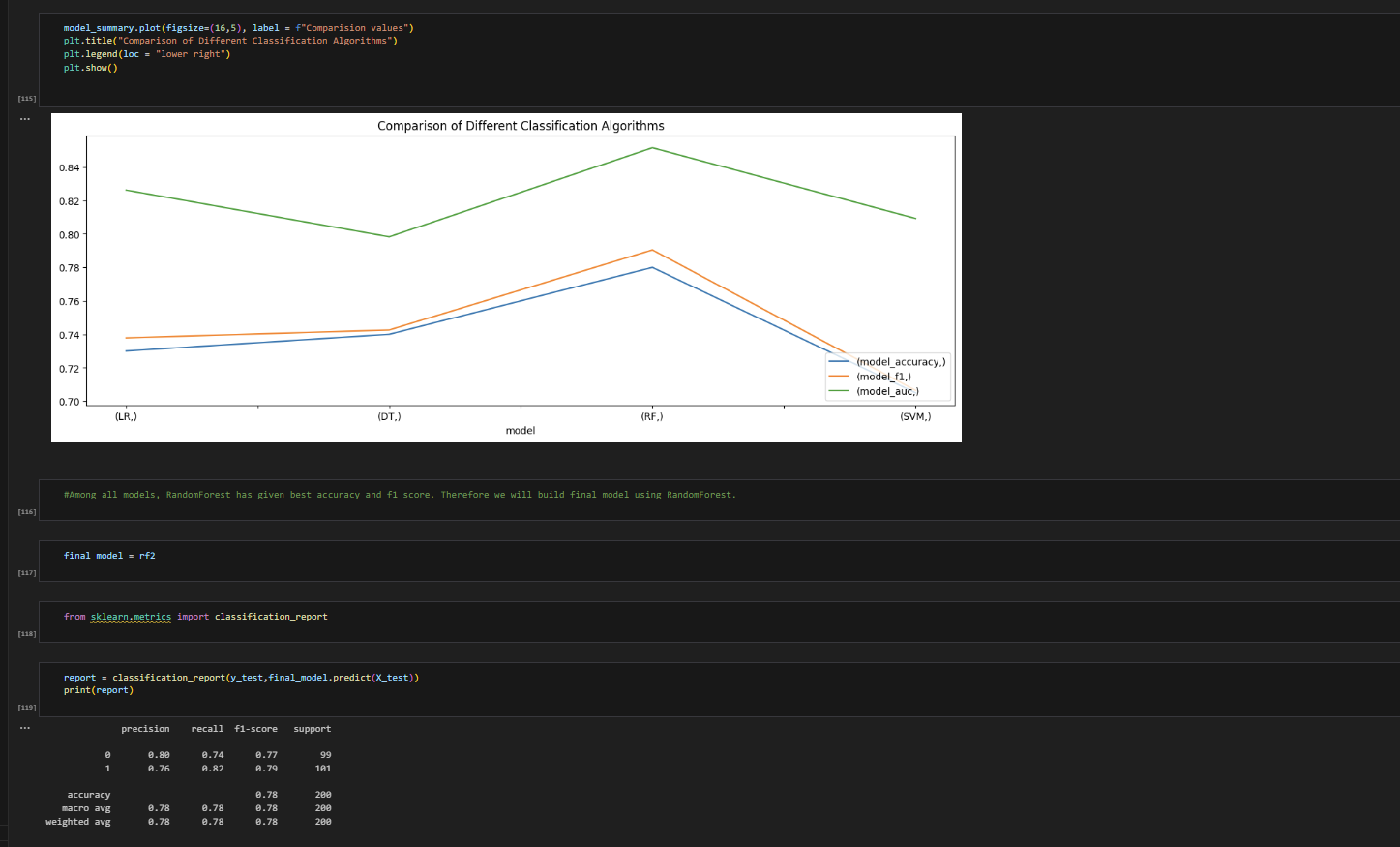
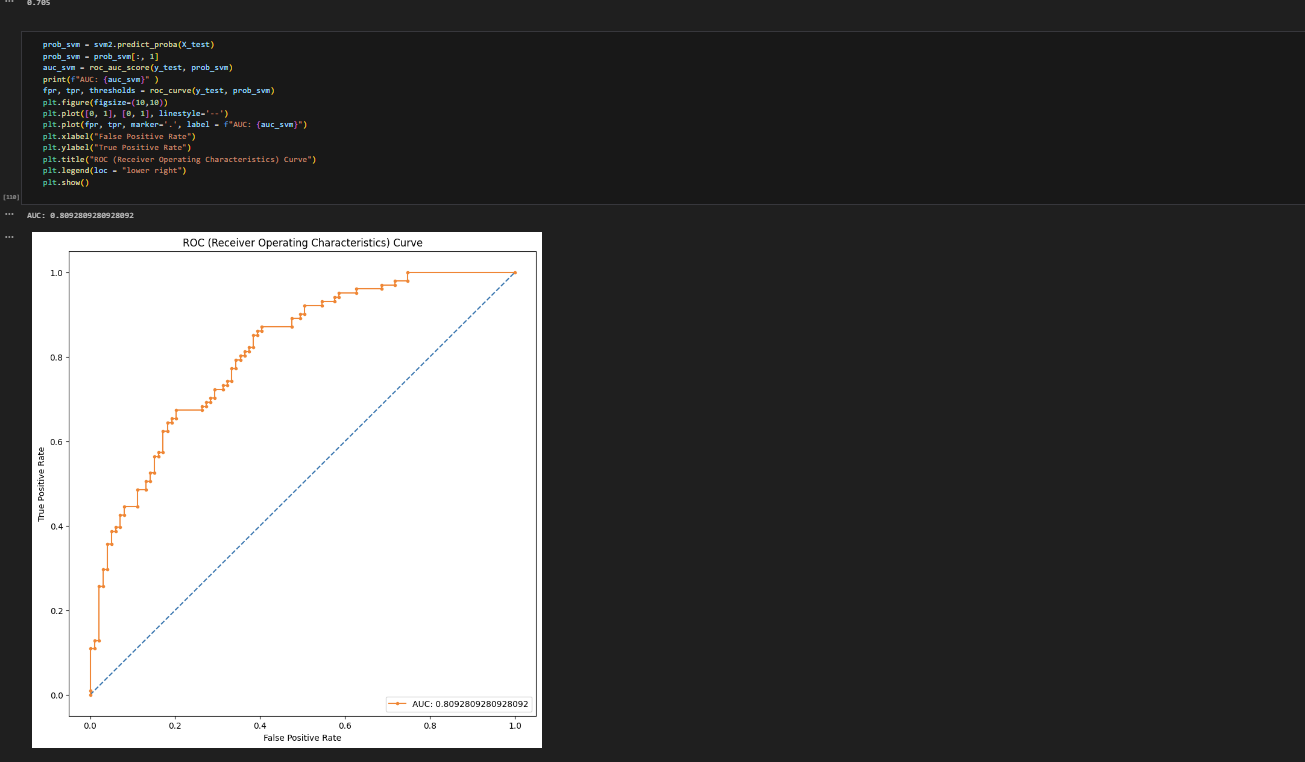
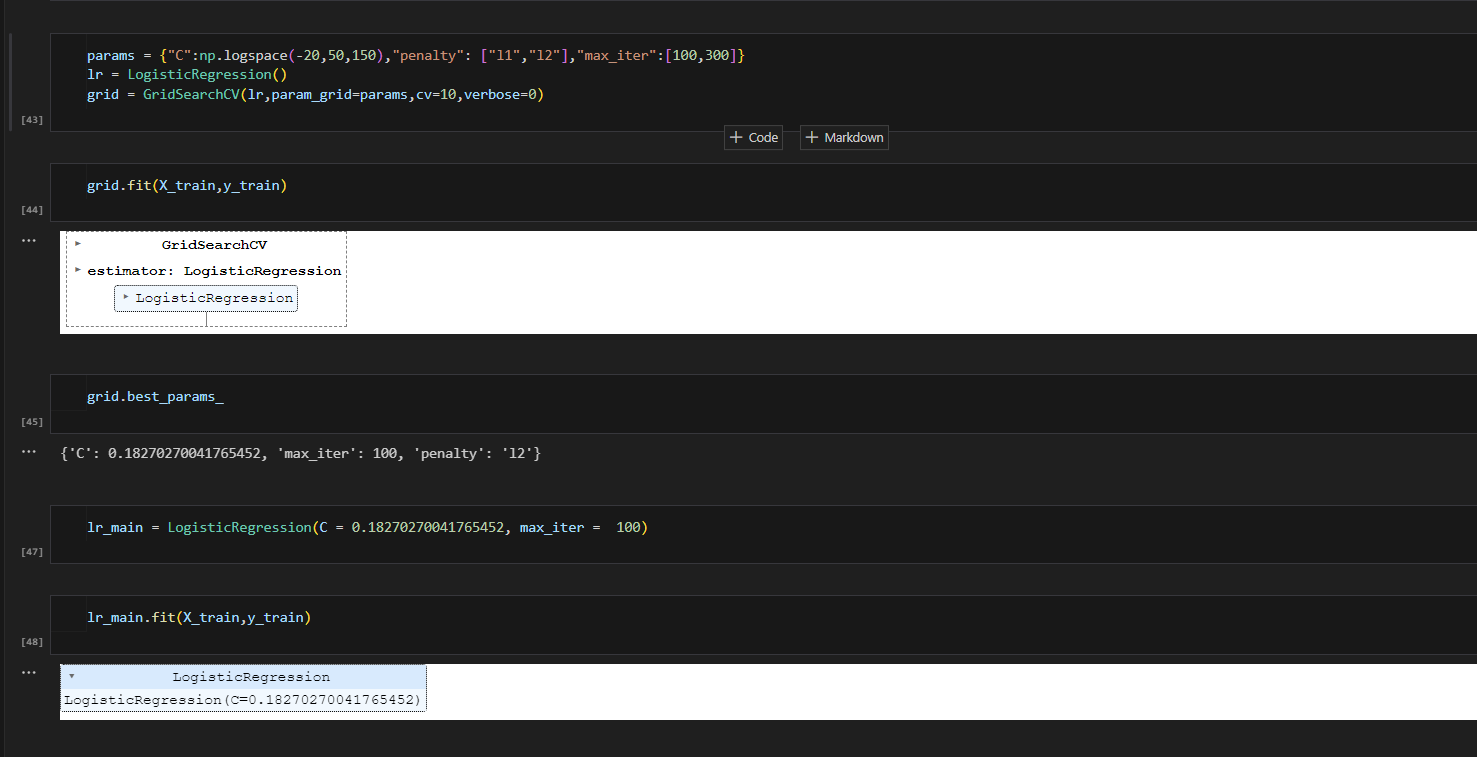
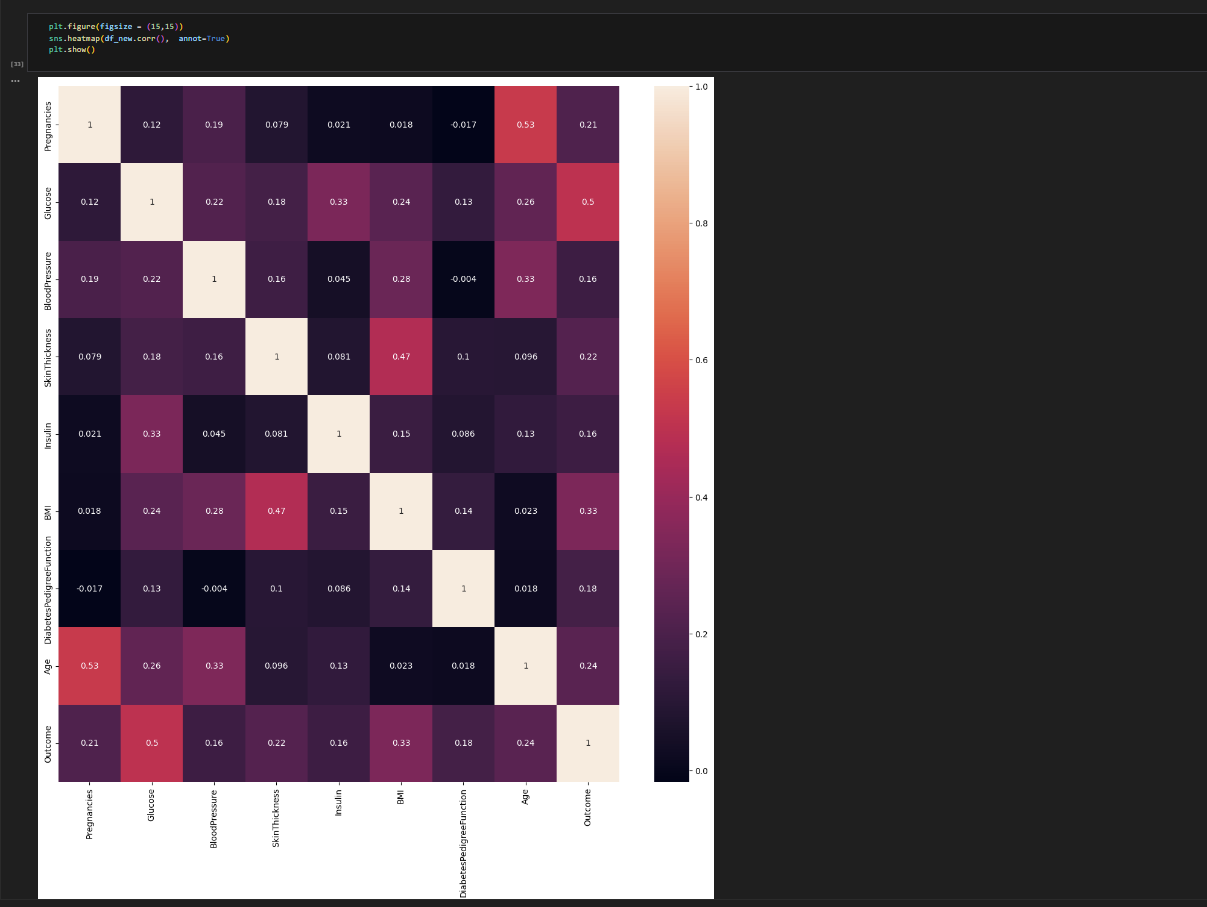
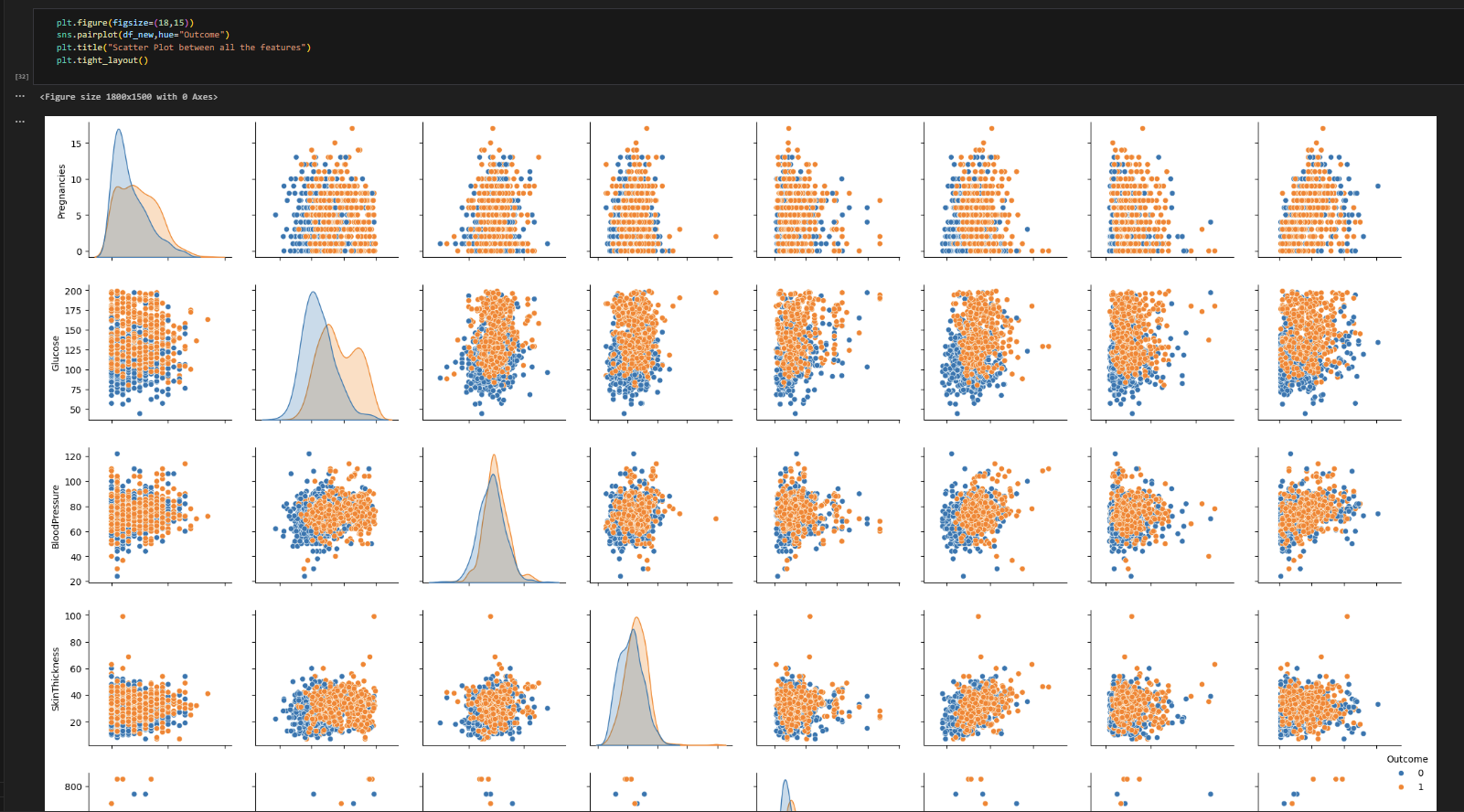
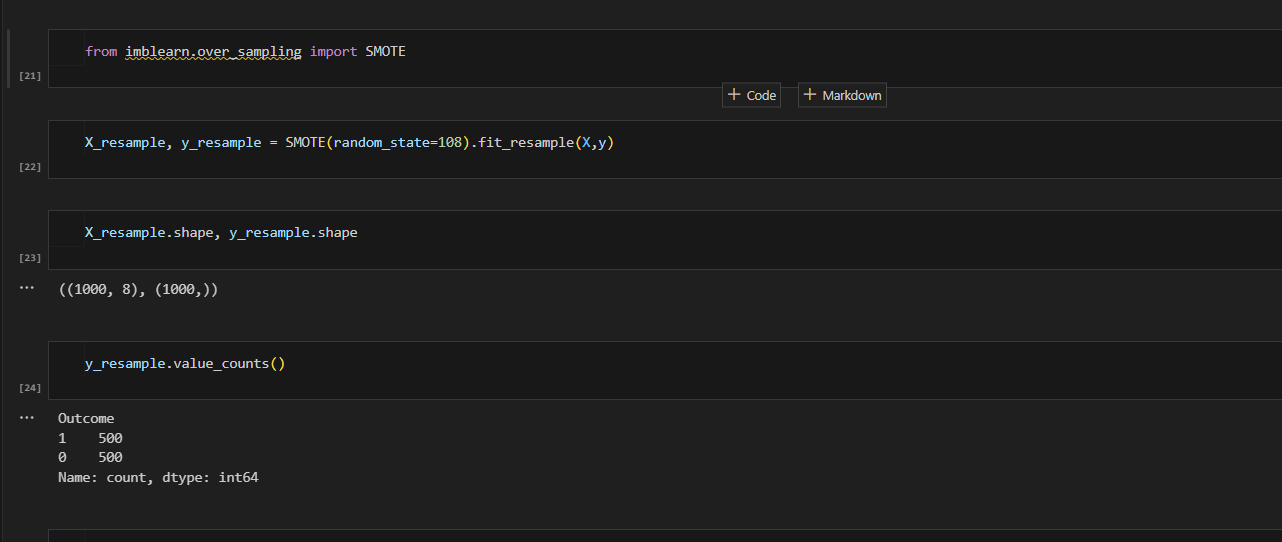
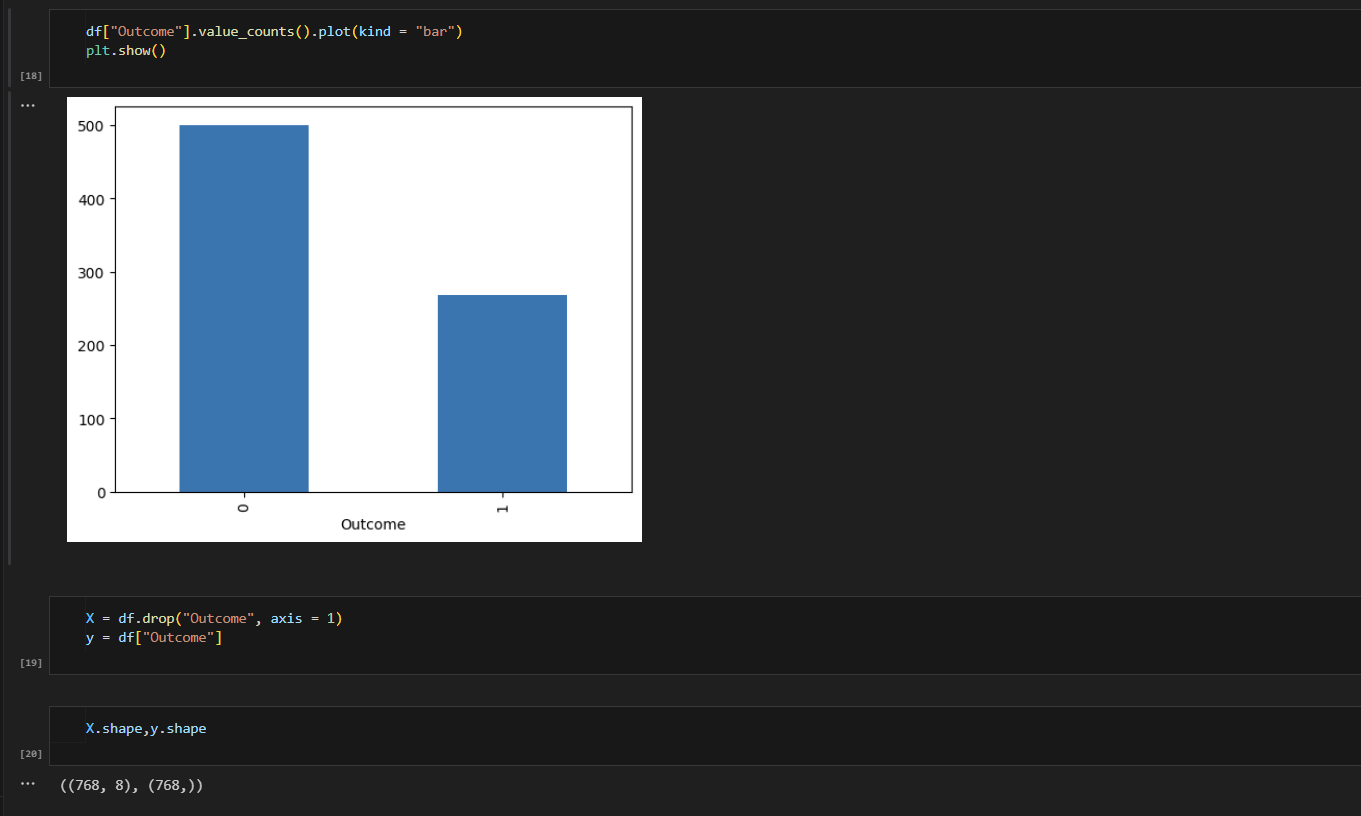
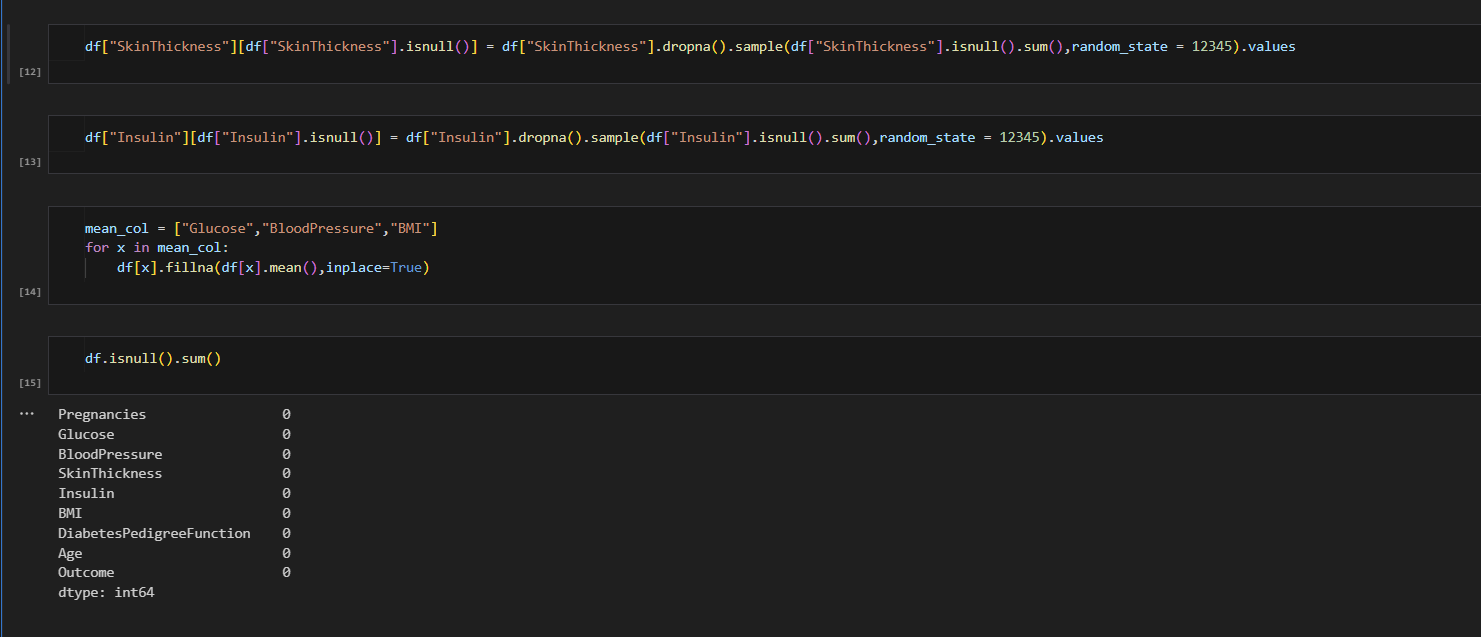
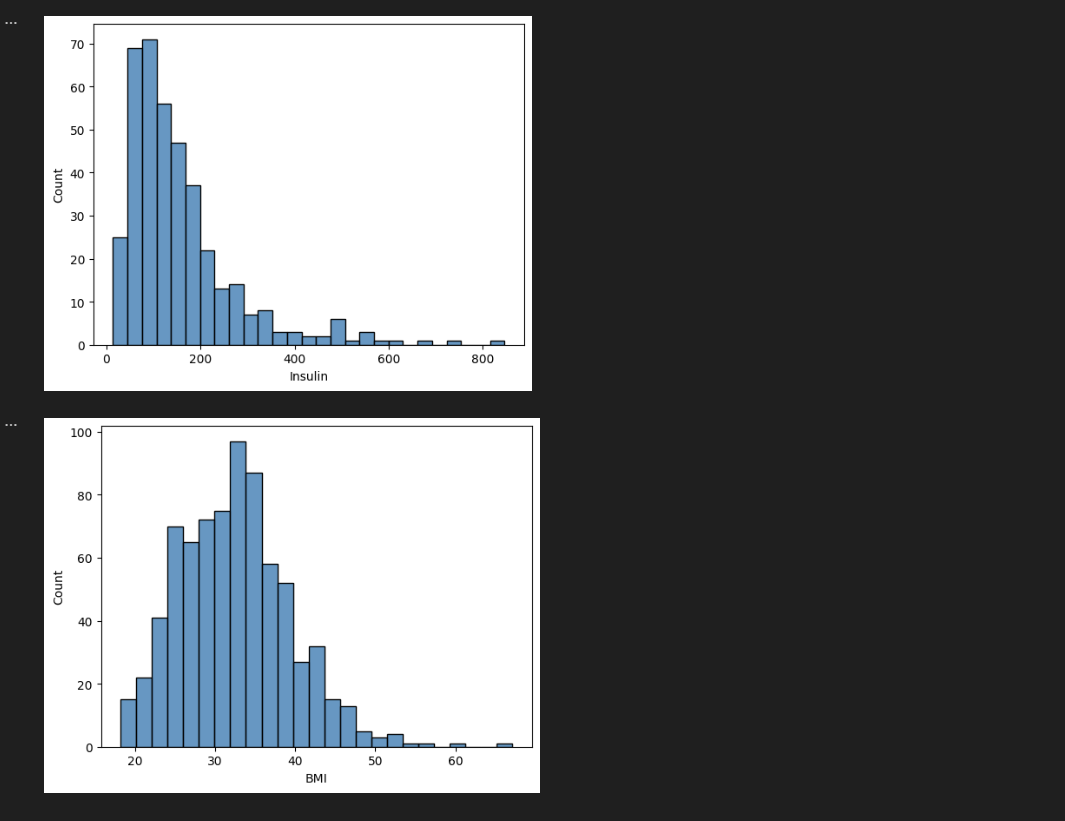
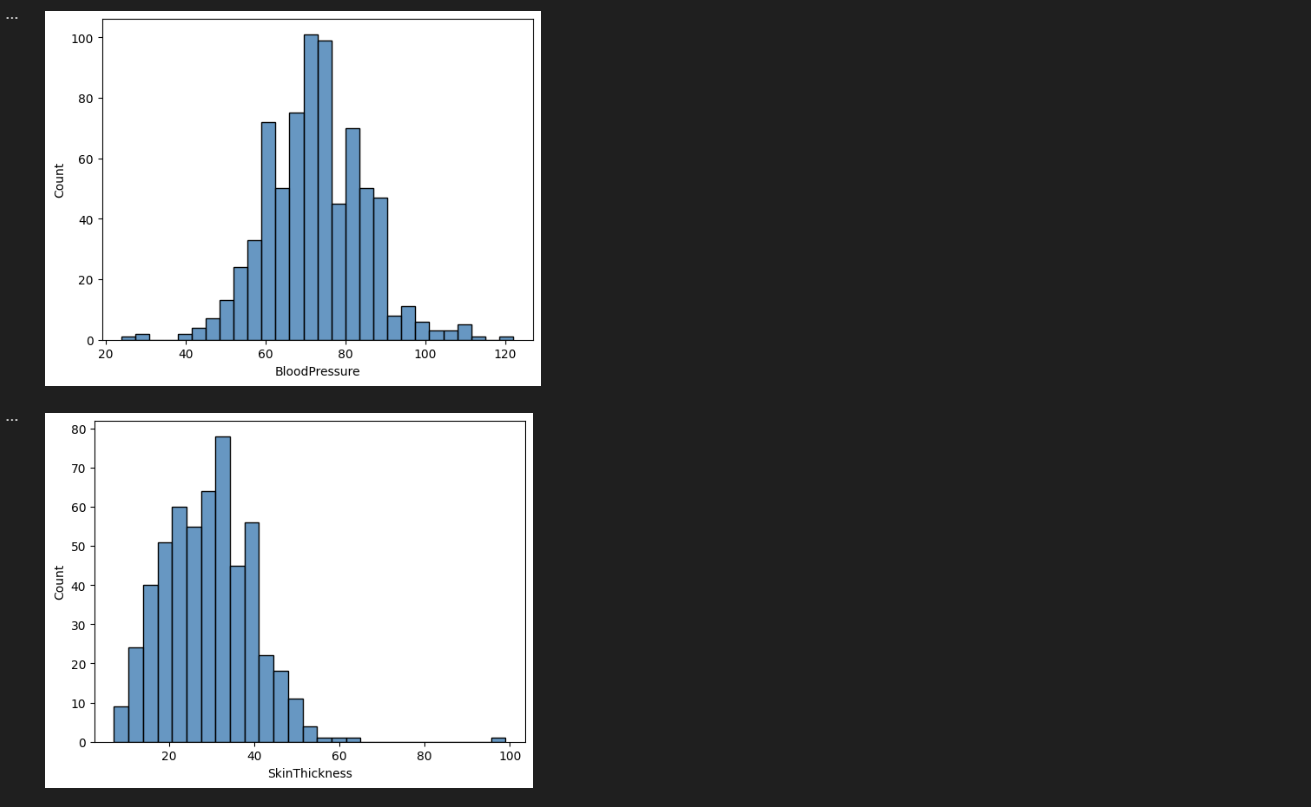
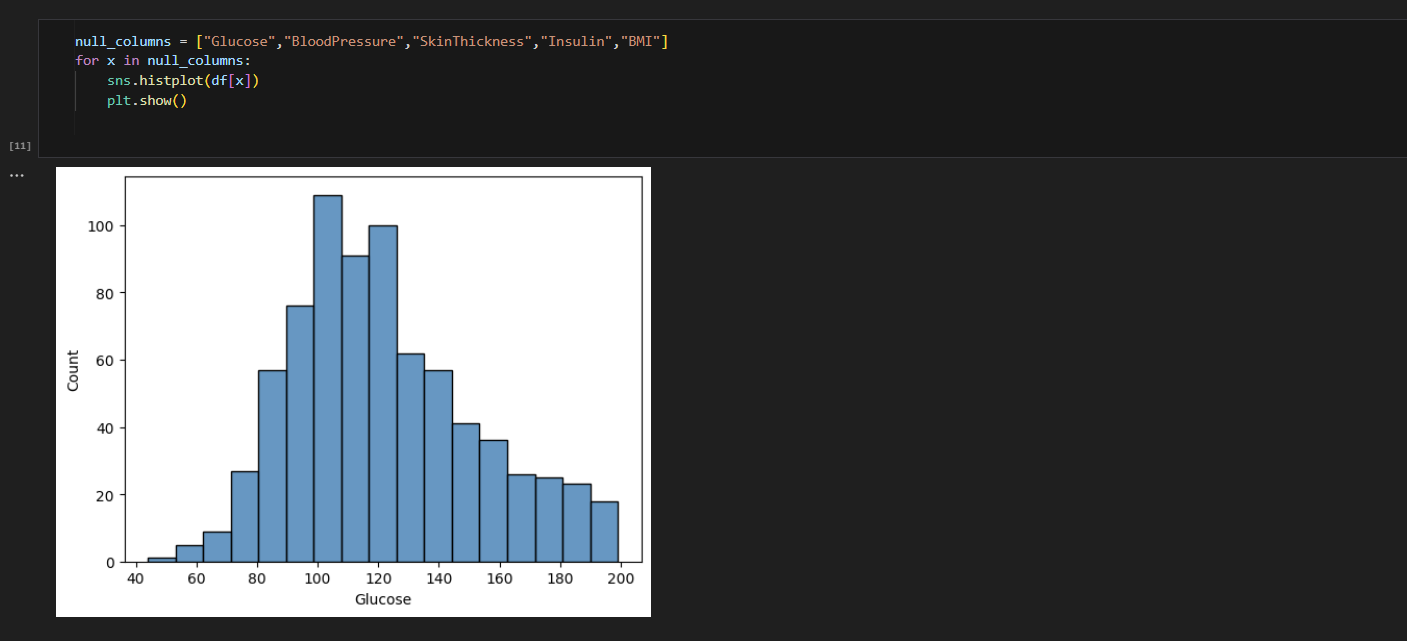
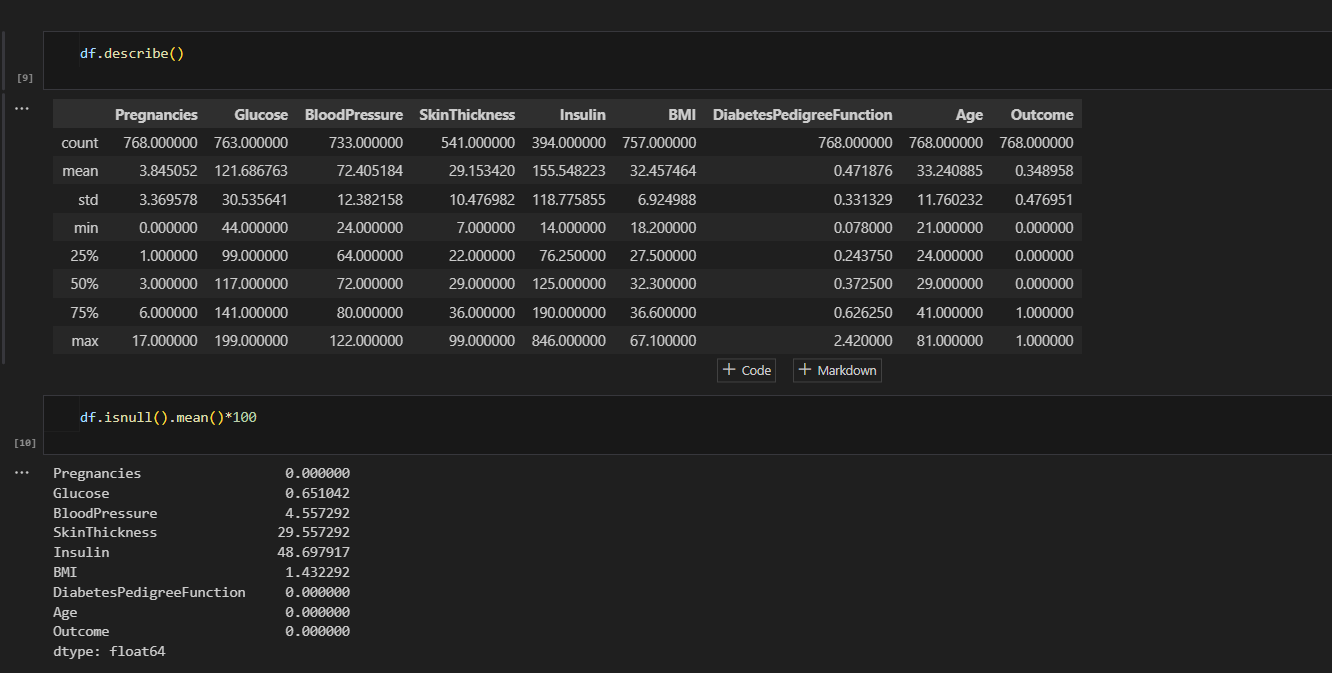
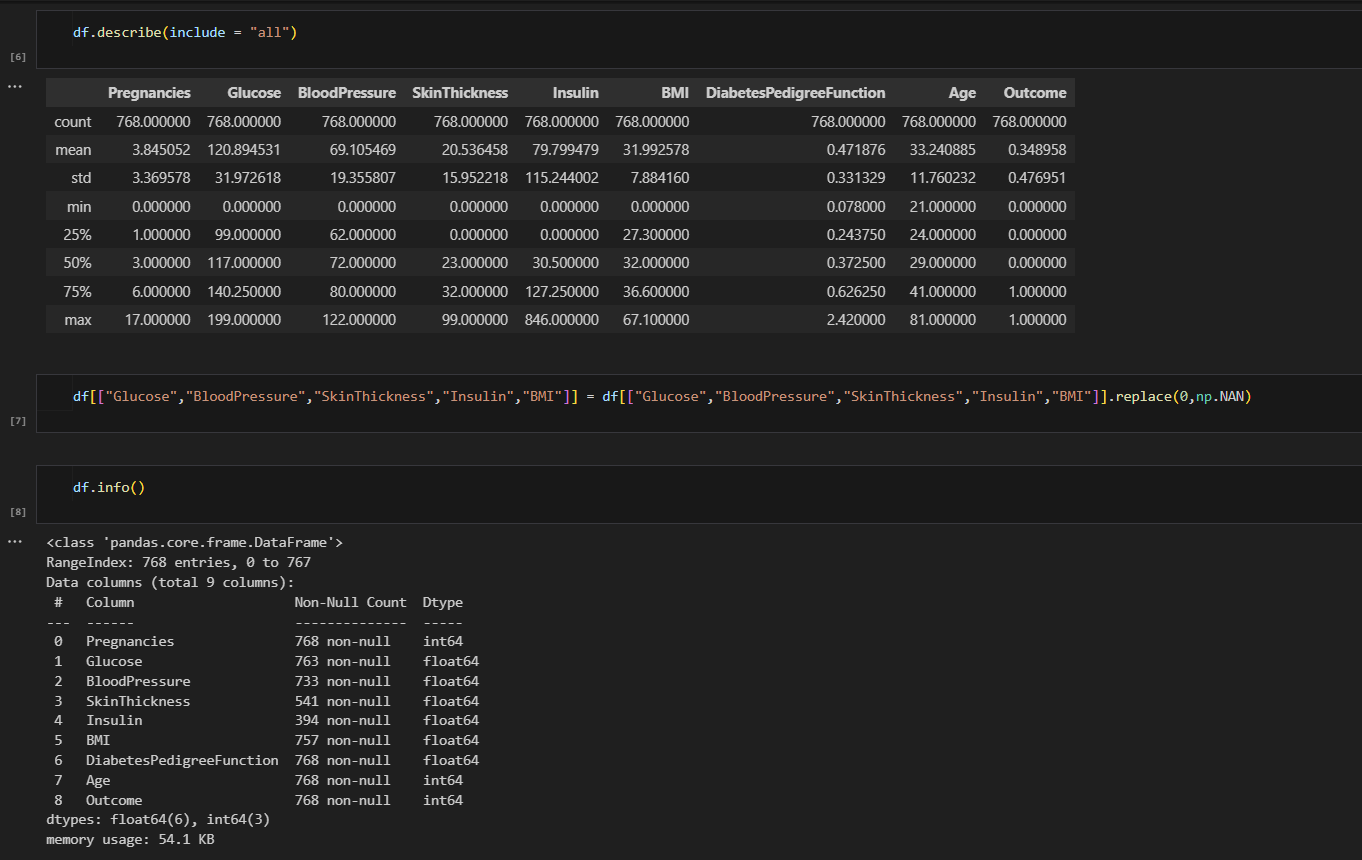
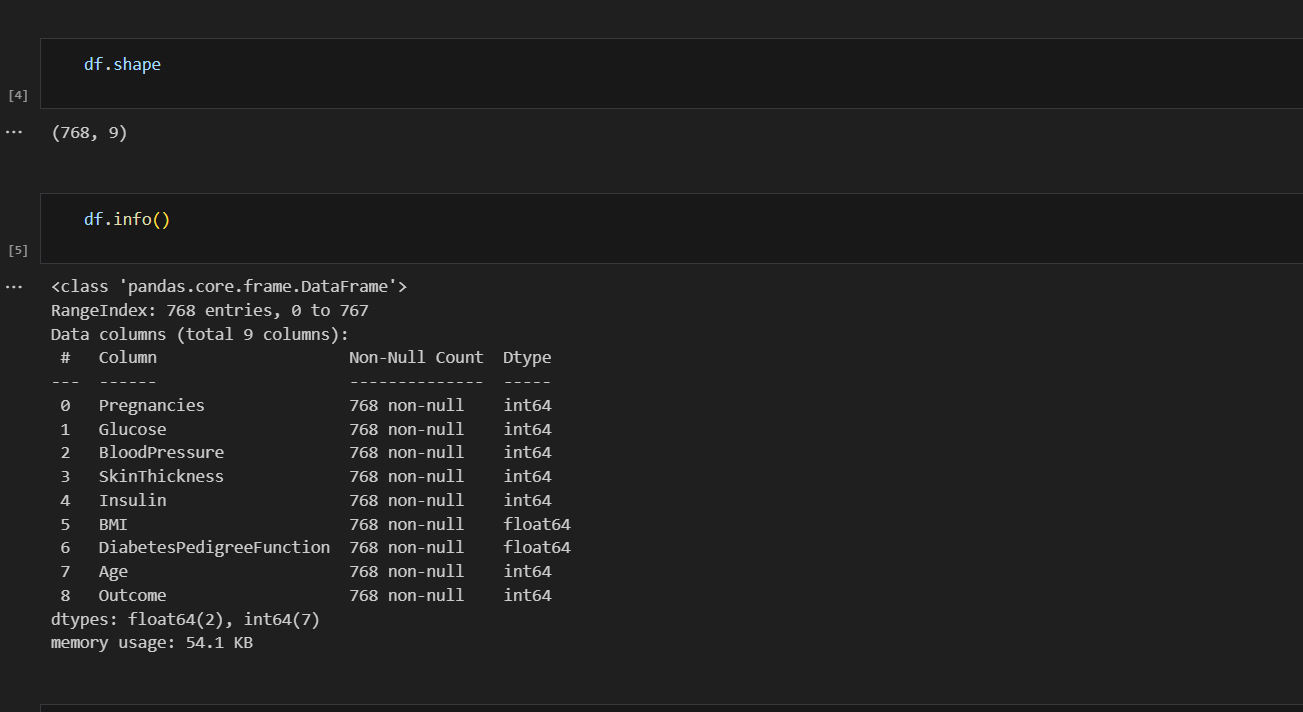
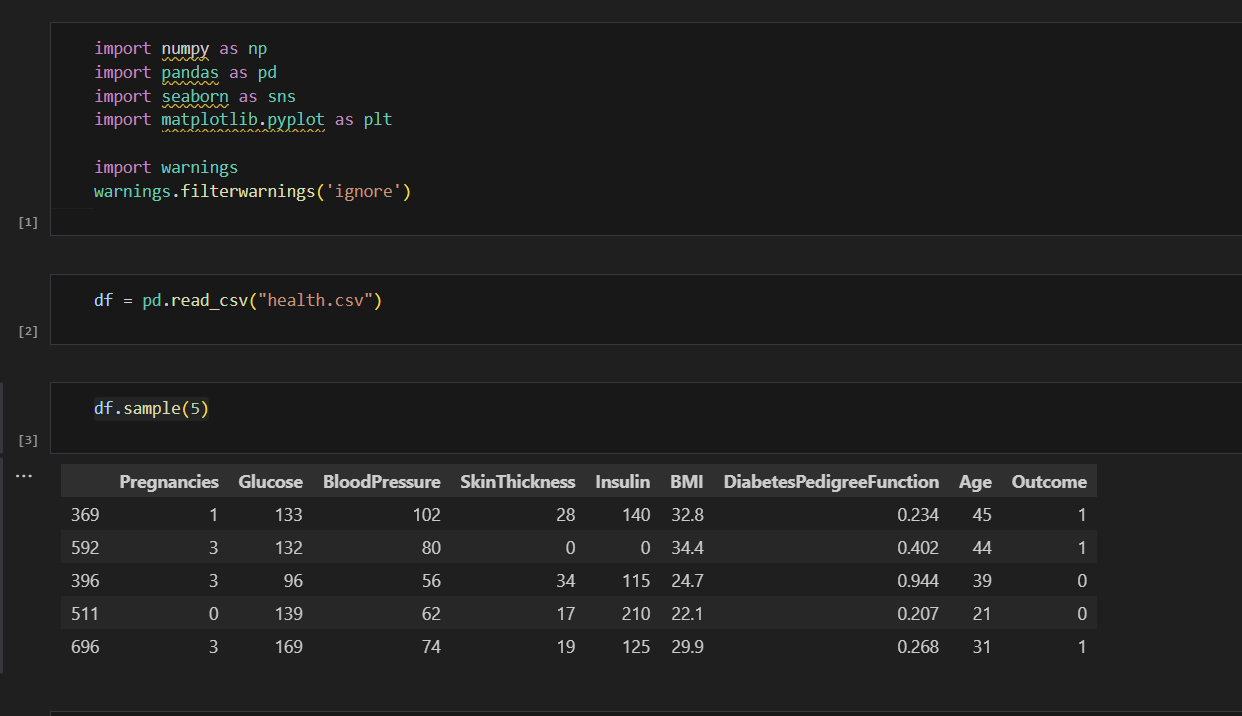
ensemble techniques to predict the diabetes dataset. We have applied SVM, LR, DT and RF

Machine learning classifier to analyse the performance by finding the accuracy of each classifier

All the classifiers are implemented using scikit learn libraries in Python. The implemented

classification algorithms are described in the next section.

**IMPLEMENTATION & OUTPUT:**

****

**TEST RESULTS ANALYSIS:**

In our project, we focused on two main goals:

1. Investigate the data to find any possible links between different factors and diabetes.

2. Achieve the highest accuracy in predicting diabetes using various supervised learning algorithms.

For the first goal, our analysis suggests a positive link between high glucose levels and diabetes, but we cannot confirm a direct cause-effect relationship. For the second goal, we compared different machine learning methods and found that Random Forest produced the most accurate predictions.

Our project aimed to develop a reliable model for predicting diabetes in patients. We successfully tested a range of machine learning methods, including SVM, KNN, Random Forest, Decision Tree, Logistic Regression, and Gradient Boosting.

The test accuracy of our models ranged from about 73% to 81%. Overall, the Random Forest Classifier provided the most accurate results based on both accuracy and recall scores.

**CONCLUSION:**

The project aimed to create a model to identify patients with diabetes who face a high risk of hospitalization. Predicting hospital admission risk is challenging due to the complexity and multifaceted nature of the factors involved. There is a pressing need for tools that can aid healthcare organizations in understanding key aspects of predicting hospital admission risk for diabetes patients. This project contributes to existing diabetes detection methods by offering a system that assists in identifying patients who have a higher risk of developing diabetes.

The project uses machine learning techniques to analyze critical factors such as blood glucose levels, body mass index, and other medical records, enabling retrospective analysis of patients' health data. The model predicts a person's risk of developing diabetes based on medical information provided through an online web application. Users input relevant medical data, which is then processed by the trained model to predict whether the individual is diabetic or not.

The model is built using an artificial neural network comprising six dense layers, each playing a vital role in its overall performance. The model demonstrates a strong accuracy rate of 98%, indicating its effectiveness and dependability in making predictions.

**References:**

[1] Journal of Computational Intelligence in Bioinformatics ISSN 0973-385X Volume 10,

Number 1 (2017) pp. 1-8 © Research Foundation http://www.rfgindia.com

[2]A Method for Classification Using Machine Learning Technique for Diabetes Aishwarya.

School of Computing Science and Engineering, VIT University, Vellore – 632014, Tamil

Nadu, India.

[3]2) Causes of vision loss worldwide, 1990-2010: a systematic analysis.

Bourne RR, Stevens GA, White RA, Smith JL, Flaxman SR, Price H et al. Lancet Global

Health 2013;1: e339-e349

[4] A. Nather, C.S. Bee, C.Y. Huak, J.L.L. Chew, C.B. Lin, S. Neo, E.Y. Sim,

Epidemiology of diabetic foot problems and predictive factors for limb loss,

J. Diab. Complic. 22 (2) (2008) 77–82.

[5] Shiliang Sun, A survey of multi-view machine learning, Neural Comput. Applic.

23 7–8) (2013) 2031–2038.

[6]R.Alizadehsani, M.H. Zangooei, M.J. Hosseini, J. Habibi, A. Khosravi, M. Roshanzamir,

F. Khozeimeh, N. Sarrafzadegan, S. Nahavandi, Coronary artery disease detection using

computational intelligence methods, Knowledge-Based Systems, 109 (2016) 187-197.

[7] P. Sattigeri, J.J. Thiagarajan, M. Shah, K.N. Ramamurthy, A. Spanias, A scalable

feature learning and tag prediction framework for natural environment sounds, Signals Syst.

and Computers 48th Asilomar Conference on Signals, Systems and Computers, 2014, 1779

1783.