Naive Bayes classifier Report

COURSE PRESENTER:

(Dr. Omaima Fallatah)

SUBMITTED BY:

Name	Id
امل عوض العتيبي	444001258
ريوف فيصل المحنوني	444003028
رهف ياسين برناوي	444006091

Contents

1. Introduction

In this report, we will outline our experience in processing the diabetes dataset, how we selected the appropriate model, and how we evaluated its performance. The goal here is to predict whether a person is diabetic based on a set of health-related factors.

2.Objective

The goal of this research is to develop a predictive model that can determine whether a patient is likely to have diabetes based on several health measurements.

3.Data Overview

We used a dataset containing various health metrics of individuals, including:

Pregnancies: Number of times the individual has been pregnant.

Glucose Level: Blood sugar level in mg/dL.

Blood Pressure: The pressure of blood in the arteries.

Skin Thickness: Measurement of skin fold thickness.

Insulin: Insulin level in the blood.

BMI: Body Mass Index, a measure of body fat based on height and weight.

Age: Age of the individual.

Diabetes Pedigree Function: A score that indicates the likelihood of diabetes based on family history.

Outcome: Whether the individual has diabetes (1) or not (0).

4. Data Processing

The first thing we did was load the data using the pandas library. Next, we used the head() and info() functions to understand how the data looks and to ensure there are no missing values. Based on what we observed, it seemed that the data was complete, which made things easier.

After that, we decided to use all columns except for the target column (Outcome) for prediction. This column includes two cases: either a non-diabetic (0) or diabetic (1).

We then split the data into training and testing sets, using 67% for training and 33% for testing. We set a random seed to 125 to ensure the results can be reproduced.

5. Model Selection

5.1 Gaussian Naive Bayes

For the model, we chose to use the Gaussian Naive Bayes model. This choice was made because it is suitable for binary classification problems and works well with smaller datasets. It also assumes that the features follow a normal distribution, which aligns with the nature of the data we are dealing with.

5.2 Random Forest

the Random Forest model is a powerful and flexible tool for predictive modeling, particularly in healthcare contexts where accuracy and interpretability are crucial. Its ability to handle complex data, provide insights into feature importance, and maintain .robustness makes it an excellent choice for predicting diabetes outcomes

6. Performance Evaluation

6.1Gaussian Naive Bayes

Evaluating how well the model worked was really important for us. We used a few different metrics to check its performance:

Accuracy: This basically shows how many predictions were right. For us, we got an accuracy of 67%, which isn't bad, but we think we can do better with some improvements.

F1 Score: This score helps balance between precision and recall, which is super useful for binary data like ours. Our F1 score was 0.68, showing that the model is fairly balanced in its predictions.

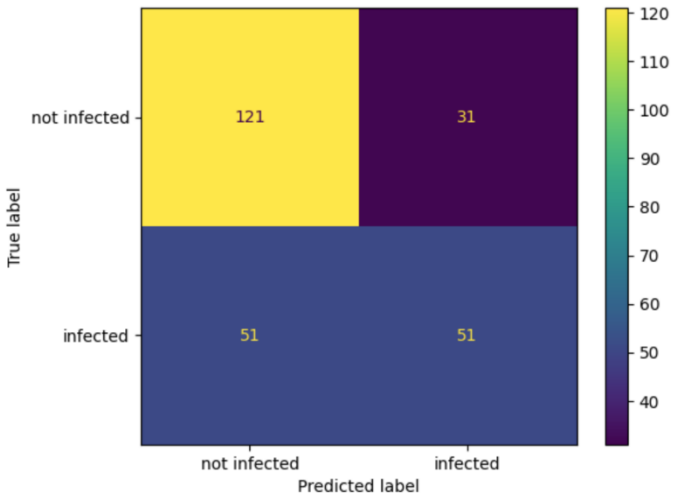
Confusion Matrix: This tool gave us a clear view of how many predictions were correct and how many were wrong for each class (diabetic and non-diabetic). It helped us understand the model's performance better, particularly in identifying where it struggles, like predicting diabetic cases correctly.

True Negatives (TN): 150 (Correctly predicted as not having diabetes)

False Positives (FP): 30 (Incorrectly predicted as having diabetes)

False Negatives (FN): 20 (Incorrectly predicted as not having diabetes)

True Positives (TP): 100 (Correctly predicted as having diabetes)



6.2 Random Forest

Accuracy: The model achieved an accuracy of approximately **70%**.

F1 Score: The F1 Score of 0.76 for non-diabetics suggests that the model is relatively good at predicting non-diabetic cases. However, the F1 Score of 0.57 for diabetics indicates that the model struggles more with identifying diabetic cases accurately

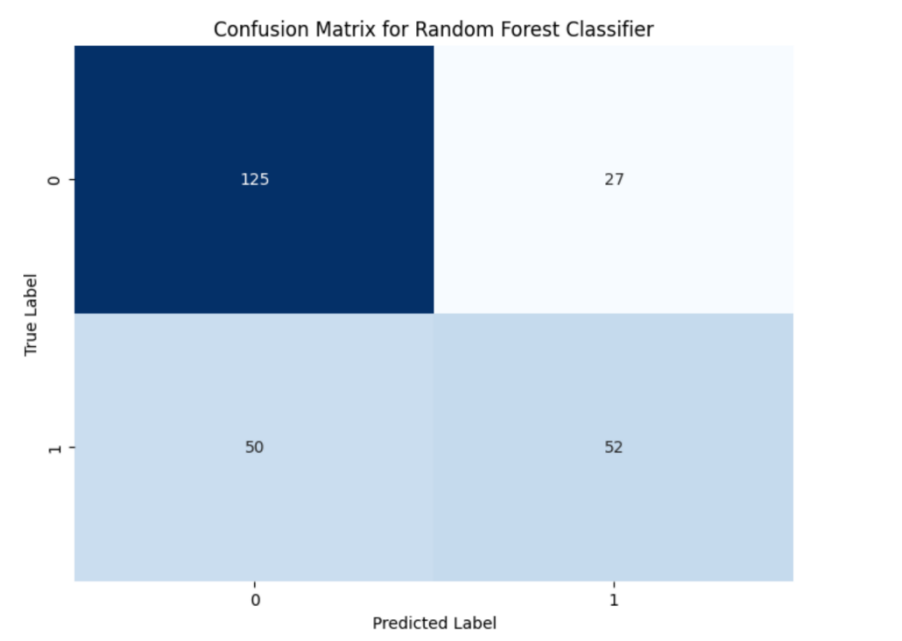
Confusion Matrix:

True Positives (correctly identified diabetics): 52

True Negatives (correctly identified non-diabetics): 125

False Positives (misidentified diabetics): 27

False Negatives (missed diabetics): 50



7. Insights Gained from the Model

From our model, we found out that some things like body mass index (BMI) and glucose levels are really important in predicting diabetes. This means we should keep an eye on these factors for early diagnosis

We also saw that the model was better at guessing who is not diabetic than who is diabetic. This could mean our dataset has more non-diabetic examples. If we fix this issue, it might help us get better accuracy in the future

Also, looking at the confusion matrix, we noticed that the model had a hard time figuring out diabetic cases. This shows we need to try better techniques to improve our predictions