NAIVE BAYES CLASSIFICATION

By students :

Hala faiz alharbi	444001733
Rawyah Alasmari	443005517

introduction

In this report, we will analyze heart disease data and explain the models used. Our goal is to determine whether the patient is suffering from heart disease or not.

Objective

Our goal is to determine whether the patient is suffering from heart disease using the appropriate model.

Data Overview

The data we have was taken from a website called Kakel, containing 14 columns and 1025 rows. The columns we have are:

- age: age of the patient
- sex: sex of the patient
- cp: chest pain type
- trestbps: resting blood pressure
- chol: serum cholesterol
- fbs: fasting blood sugar
- restecg: resting electrocardiographic results
- thalach: maximum heart rate achieved
- exang: exercise induced angina
- oldpeak: ST depression induced by exercise relative to rest
- slope: slope of the peak exercise ST segment
- ca: number of major vessels colored by fluoroscopy
- thal: thalassemia
- target: the outcome indicating whether the patient has heart disease or not.

The target is integer valued: 0 = no disease and 1 = disease.

Data Preprocessing

First, the data was downloaded from a website, then uploaded to a Google Colab page using the Pandas library. We read the file and used df.head() and df.shape to display the data and understand it, as well as to know how many rows and columns the dataset contains. Then, we checked the cleanliness of the data and ensured there were no missing or anomalous values that could affect the results by using the command missing_values = df.isnull().sum(). The result showed that there were no missing values, confirming that the data is ready for analysis.

After ensuring that the data is ready for analysis, we decided to use all columns except for the target column. This column includes two cases: either a person is not infected (0) or infected (1). We then divided the data into training and testing sets, using 80% for training and 20% for testing.

Model Selection

1- Gaussian Naive Bayes

We decided to use the Gaussian Naive Bayes model. This option is suitable for binary classification tasks, as it shows good efficiency even with limited data and is also fast in training and prediction.

2- Multinomial Naive Bayes

We decided to use the Multinomial Naive Bayes model. This option is suitable for classification tasks, especially in cases where the data represents multiple features. This model is highly efficient in handling large datasets and is fast in both training and prediction, making it particularly suitable for predicting whether a person has a disease or not.

3- DecisionTree model

We decided to use the Decision Tree model. This option is ideal for classification tasks because it allows for clear interpretation of results. The model is capable of handling complex data and works well with both categorical and continuous features. Additionally, it can uncover patterns and relationships between variables, making it suitable for accurate predictions.

Performance Evaluation

1- Gaussian Naive Bayes

- Accuracy: This primarily reflects the percentage of correct predictions. For us, we achieved an accuracy of 80%, which is a good result that reflects the model's performance.
- Precision: We had a precision of 0.87 for the "not infected" (0) class and 0.75 for the "infected" (1) class. This indicates that the model achieves a high rate of correct predictions for the non-infected class.
- Recall: The model's recall was 0.71 for the "not infected" class and 0.89 for the "infected" class, demonstrating the model's ability to identify positive cases effectively.
- F1 Score: The F1 score was 0.78 for the "not infected" class and 0.82 for the "infected" class, indicating a good balance between precision and recall.
- Data Summary: A total of 205 cases were used in the analysis, with the following results:
- True Negatives (TN): 102 (correctly predicted not infected).
- False Positives (FP): 30 (incorrectly predicted not infected).
- False Negatives (FN): 20 (incorrectly predicted infected).
- True Positives (TP): 103 (correctly predicted infected).

2- Multinomial Naive Bayes

- Accuracy: This primarily reflects the proportion of correct predictions. For us, we achieved an accuracy of 68.78%, which is a reasonable result but can be improved.
- Precision: We had a precision of 0.70 for the "not infected" (0) class and 0.68 for the "infected" (1) class, indicating that the model achieves a good rate of correct predictions.
- Recall: The model's recall was 0.65 for the "not infected" class and 0.73 for the "infected" class, showing the model's ability to recognize positive cases at an acceptable level.
- F1 Score: The F1 score was 0.67 for the "not infected" class and 0.70 for the "infected" class, indicating a good balance between precision and recall.
- Total Data: A total of 205 cases were used in the analysis, with the following results:
- True Negatives (TN): 102 (correctly predicted not infected).
- False Positives (FP): 30 (incorrectly predicted not infected).
- False Negatives (FN): 20 (incorrectly predicted infected).
- True Positives (TP): 103 (correctly predicted infected).

3- DecisionTree model

- Accuracy: This primarily reflects the percentage of correct predictions. For us, we achieved an accuracy of 98.54%, which is an excellent result that reflects the model's performance.
- Precision: We had a precision of 0.97 for the "not infected" (0) class and 1.00 for the "infected" (1) class, indicating that the model achieves a very high percentage of correct predictions.
- Recall: The model's recall was 1.00 for the "not infected" class and 0.97 for the "infected" class, demonstrating the model's ability to accurately identify positive cases.
- F1 Score: The F1 score was 0.99 for the "not infected" class and 0.99 for the "infected" class, indicating an excellent balance between precision and recall

Insights Gained from the Model

When comparing the models, it is clear that the Decision Tree model significantly outperforms the others in terms of accuracy and F1 score, making it the optimal choice for predicting heart disease. While Gaussian Naive Bayes is also a good option, it does not reach the accuracy level of the Decision Tree. On the other hand, Multinomial Naive Bayes requires significant improvements to be effective in this context. To enhance performance, other types of models can be used for further improvement. Additionally, increasing the dataset size by collecting more information can help improve model accuracy and reduce errors.