Mini Project Report

Submitted By

Name : Nensi Chavda & Darshita Bhatt
Enrolment No.: 12202040501011 & 12202040501015

Course Code: 202046702

Course Name: Artificial Intelligence and Machine Learning

In Partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

In

Computer Engineering

G. H. Patel College of Engineering & Technology

The Charutar Vidya Mandal (CVM) University, Vallabh Vidyanagar – 388120

Title: Heart Disease Prediction Using Machine Learning

Objective:

- The primary objective of this project is to develop a predictive model that accurately classifies whether an individual is at risk of heart disease based on health indicators. The goal is to compare multiple machine learning algorithms to identify the most effective approach, offering insights into early detection and preventive care for heart-related conditions.

Dataset Used:

- Source: Kaggle (Heart Disease Health Indicators Dataset - BRFSS2015)
- File Name: heart_disease_health_indicators_BRFSS2015.csv
- Size: 20,000 rows and multiple health-related features such as BMI, smoking status, stroke history, physical activity, diabetic status, etc.
- Label Column: HeartDiseaseorAttack (Binary: 0 - No, 1 - Yes)

Models Chosen and Training Flow:

This project follows a layered approach by implementing and evaluating three widely used classification models individually:

- **Logistic Regression**
  - Linear model for binary classification.
  - Baseline to understand the data and get initial results.
- **Support Vector Machine (SVM)**
  - Used for margin maximization and better classification on complex boundaries.
  - Kernel trick allows nonlinear decision boundaries.

- **Random Forest Classifier**
  - Ensemble of decision trees.
  - Performs better on large datasets with better generalization and reduced overfitting.

  Each model was:

  - Trained separately on the training data.
  - Evaluated individually on the test data.
  - Compared using common evaluation metrics.

Performance Metrics Used:

To evaluate the effectiveness of each model, the following performance metrics were considered:

- Accuracy: Measures the percentage of correctly predicted observations.
- Precision: Measures how many of the positively predicted cases are actually positive.
- Recall (Sensitivity): Measures how many actual positive cases were correctly predicted.
- F1 Score: Harmonic mean of precision and recall. Balances false positives and false negatives.
- Confusion Matrix: Visual representation of true positives, false positives, true negatives, and false negatives.
- Macro and Weighted Averages: To handle class imbalance in the dataset.

For linear Regression model:

```
Accuracy: 0.9004

Confusion Matrix:
[[40803   466]
 [ 4112   576]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.91      0.99      0.95     41269
         1.0       0.55      0.12      0.20      4688

    accuracy                           0.90     45957
   macro avg       0.73      0.56      0.57     45957
weighted avg       0.87      0.90      0.87     45957
```

For Random Forest model:

```
Accuracy: 0.8948364775768654
Confusion Matrix:
 [[40602   667]
 [ 4166   522]]
Classification Report:
              precision    recall  f1-score   support

         0.0       0.91      0.98      0.94     41269
         1.0       0.44      0.11      0.18      4688

    accuracy                           0.89     45957
   macro avg       0.67      0.55      0.56     45957
weighted avg       0.86      0.89      0.87     45957
```

For SVM model:

```
Accuracy: 0.8995

Classification Report:
              precision    recall  f1-score   support

         0.0       0.90      1.00      0.95     17990
         1.0       0.00      0.00      0.00      2010

    accuracy                           0.90     20000
   macro avg       0.45      0.50      0.47     20000
weighted avg       0.81      0.90      0.85     20000
```

- ## Challenges and Learning:

  - Class Imbalance: Majority of the dataset represented healthy individuals, causing poor recall for the minority class.
  - Model Overfitting: Initial models performed well on training data but poorly on unseen data.
  - Feature Selection: Identifying and preprocessing relevant features from a large dataset.
  - Handling Warnings in Joblib: Needed to manage feature name mismatches when saving/loading models.
  - Gained hands-on experience with logistic regression, SVM, and random forest models.
  - Learned how to evaluate classification models using real-world datasets.
  - Understood the impact of class imbalance on model performance.
  - Learned how to save and load ML models using `joblib`.
  - Improved skills in using Google Colab, pandas, and scikit-learn.
  - Learned how to structure a proper machine learning pipeline, from data preprocessing to final model comparison.