Addis Ababa Institute of Technology

School of Information Technology and Engineering

Software Stream

Name	ID Number
Simon Dereje	UGR/0952/14
Lemi Dinku	UGR/3860/14
Samuel Endale	UGR/9314/14

Alzheimer's Disease Prediction

Introduction

Alzheimer’s disease is a progressive neurological disorder that requires early and accurate diagnosis for effective intervention. This report examines the application of machine learning models to predict Alzheimer’s disease diagnosis using a dataset containing patient information. The models evaluated include Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbors (KNN). Among these, the Random Forest model emerged as the top performer with an accuracy of 93.02%, demonstrating the potential of ensemble learning techniques in medical diagnostics.

Problem Definition and Algorithms

The task is to predict whether a patient is diagnosed with Alzheimer's disease based on provided features. The input consists of patient attributes such as age, cognitive test scores, and medical history, while the output is a binary classification indicating the presence or absence of Alzheimer’s disease. This task is crucial for early detection and intervention.

We evaluated four machine-learning algorithms:

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. K-Nearest Neighbors (KNN)

Below are the details of each algorithm:

- **Logistic Regression:** This is a linear model widely used for binary classification tasks. It calculates the probability of a target class using the logistic function and makes predictions based on a decision threshold (commonly 0.5). Logistic Regression is effective when the relationship between features and the target is approximately linear.
- **Decision Tree:** This model builds a tree-like structure where each internal node represents a feature threshold, each branch corresponds to a decision rule, and each leaf node indicates the output class. Decision Trees are interpretable and handle both categorical and numerical data effectively.
- **Random Forest:** An ensemble method that combines multiple Decision Trees to improve predictive accuracy and control overfitting. Each tree is trained on a

random subset of the data and features, and the final prediction is obtained by averaging (for regression) or majority voting (for classification).
- **K-Nearest Neighbors (KNN):** This is a non-parametric algorithm that classifies data points based on the majority class of their k-nearest neighbors in the feature space. KNN relies heavily on the distance metric and can perform poorly with high-dimensional data due to the curse of dimensionality.

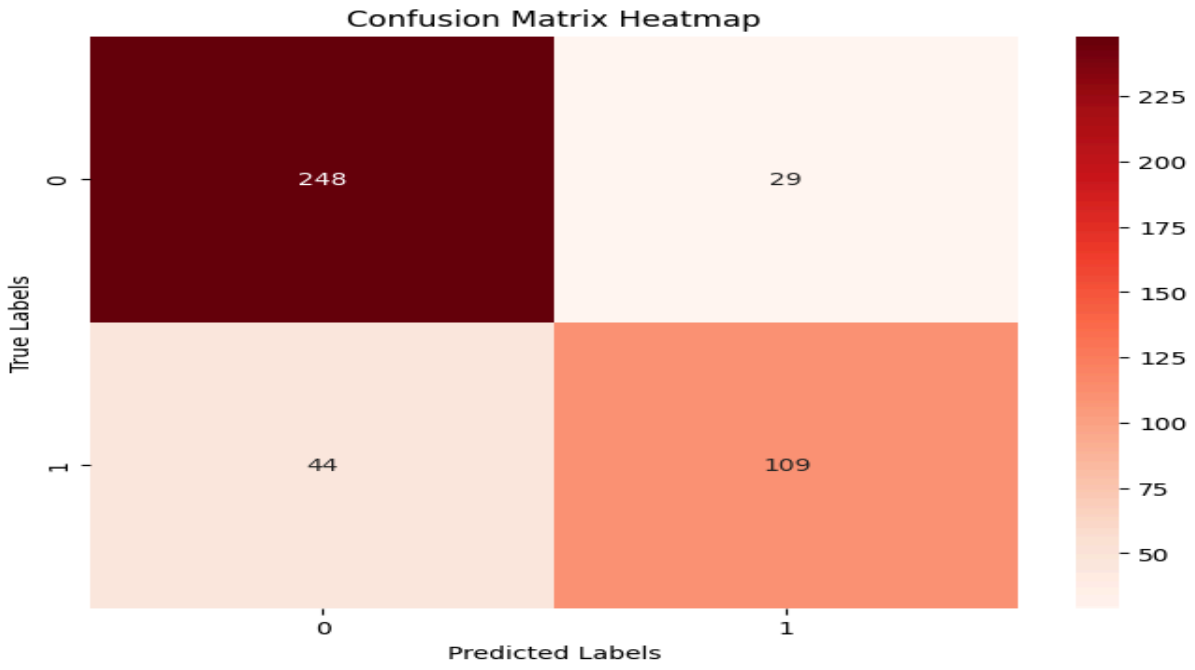
Experimental Evaluation

Methodology

- **Evaluation Criteria:** Model accuracy, precision, recall, and F1-score.
- **Hypotheses Tested:** Higher model complexity (e.g., ensemble methods) improves predictive performance.
- **Training/Test Data:** Data split into 80% training and 20% testing. StandardScaler was applied to standardize feature values.
- **Performance Metrics:** Metrics collected include accuracy, precision, recall, and F1-scores for each model.

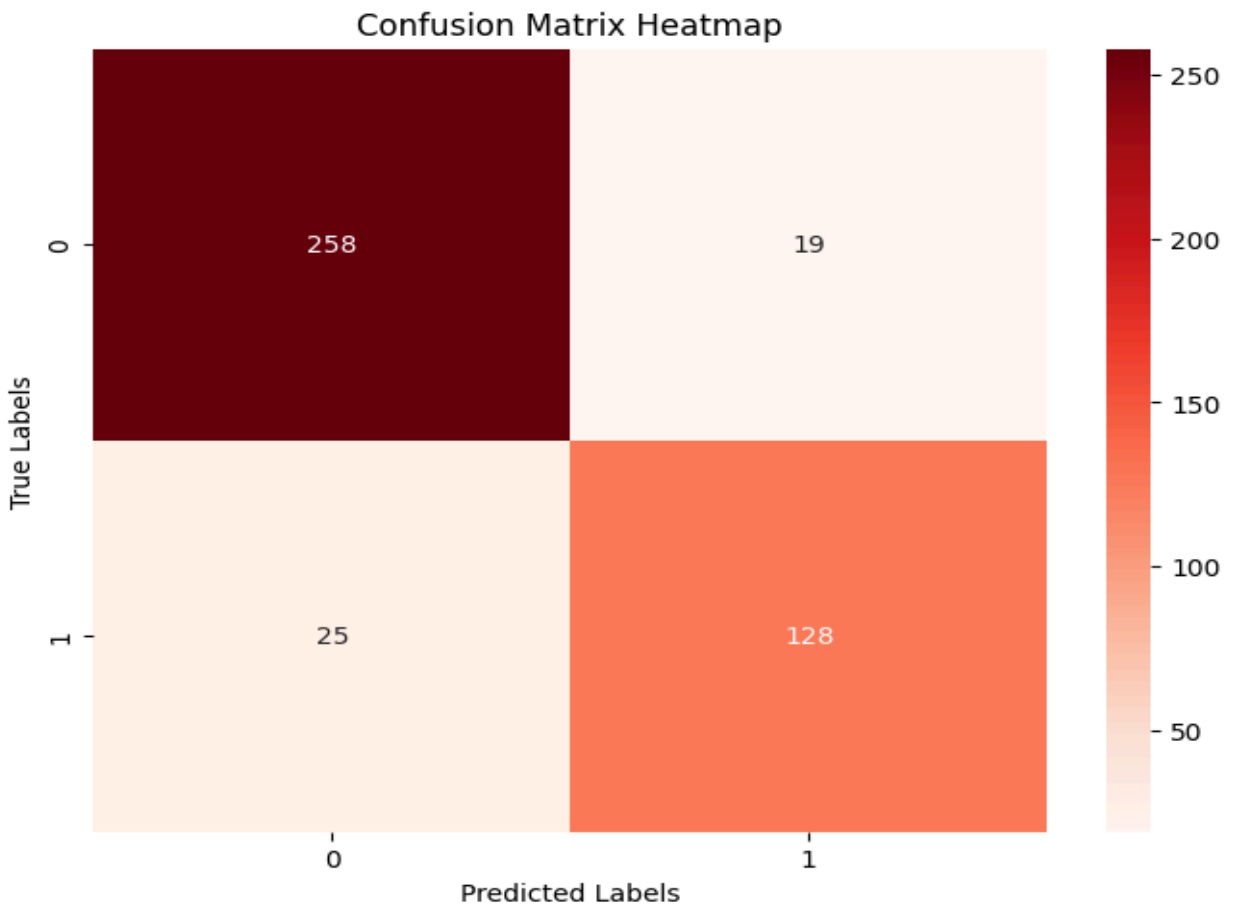
Results

1. Logistic Regression
    - **Accuracy:** 83.02%
    - **Classification Report:**
        - Precision, recall, and F1-score for class 0: 0.85, 0.90, 0.87
        - Precision, recall, and F1-score for class 1: 0.79, 0.71, 0.75
    - **Confusion Matrix Heatmap:** The confusion matrix for the Logistic Regression model shows a high number of true positives and true negatives, indicating that the model is effective at correctly classifying both classes. However, there are some false negatives, suggesting room for improvement in sensitivity.

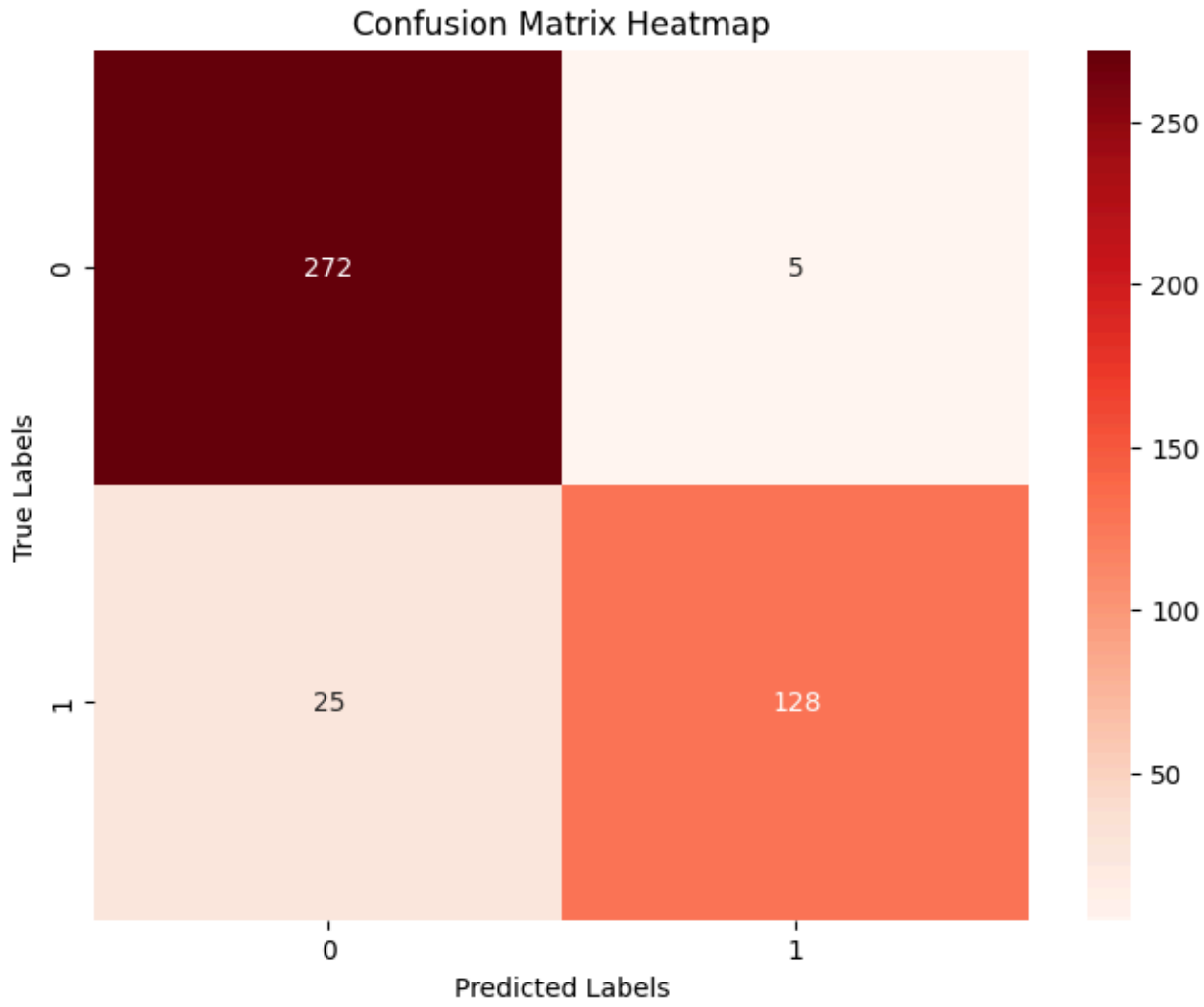


2. Decision Tree

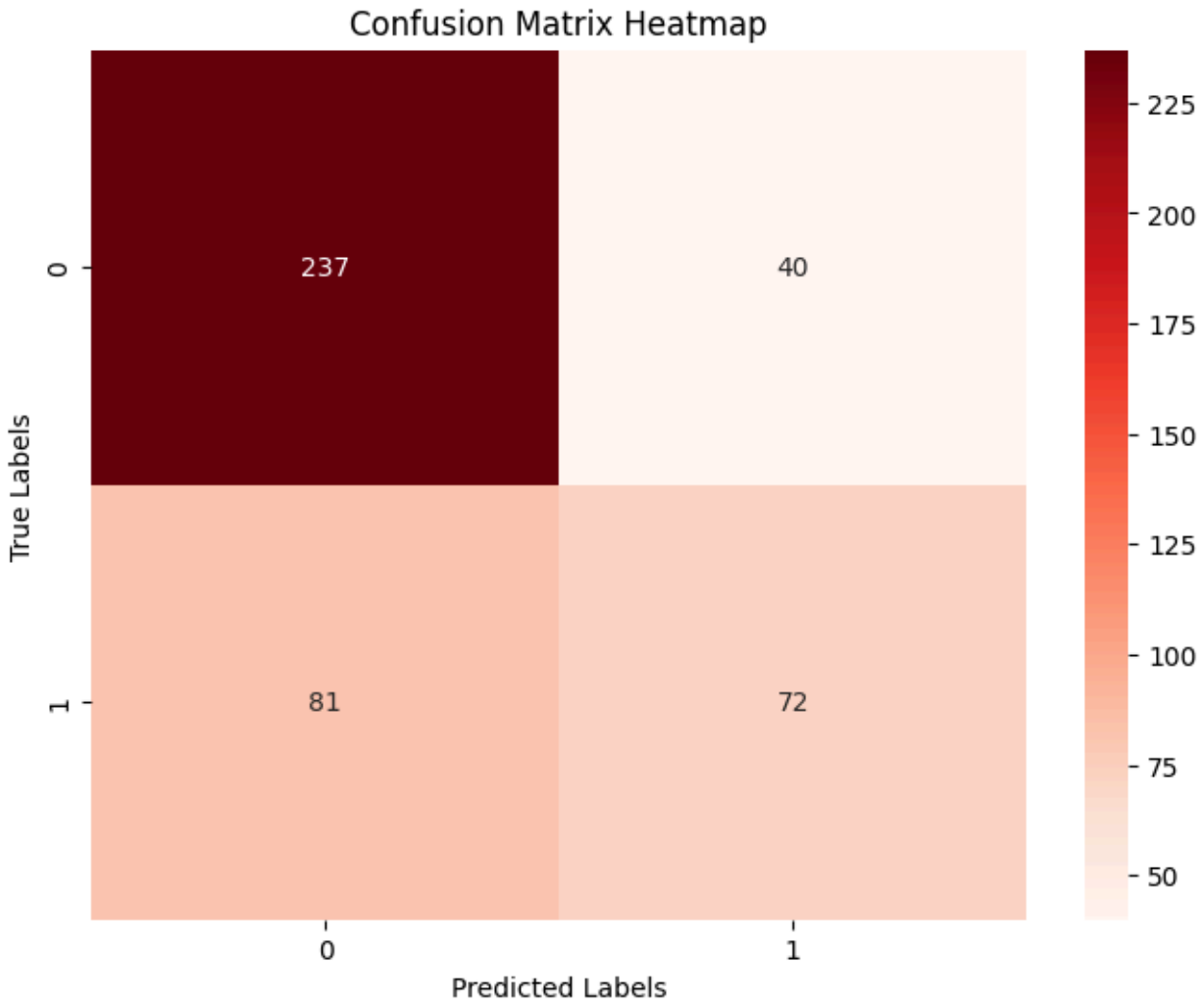
    - **Accuracy:** 89.77%
    - **Classification Report:**
        - Precision, recall, and F1-score for class 0: 0.91, 0.93, 0.92
        - Precision, recall, and F1-score for class 1: 0.87, 0.84, 0.85
    - **Confusion Matrix Heatmap:** The Decision Tree model's confusion matrix indicates a strong performance with a high number of true positives and true negatives. The model shows a balanced ability to correctly classify both classes, with fewer false positives and negatives compared to Logistic Regression.



3. Random Forest
    - **Accuracy:** 93.02%
    - **Classification Report:**
        - Precision, recall, and F1-score for class 0: 0.92, 0.98, 0.95
        - Precision, recall, and F1-score for class 1: 0.96, 0.84, 0.90
    - **Confusion Matrix Heatmap:** The Random Forest model's confusion matrix shows a very high number of true positives and true negatives, with minimal false classifications. This model typically excels in reducing overfitting and improving generalization.



4. K-Nearest Neighbors (KNN)
    - **Accuracy:** 71.86%
    - **Classification Report:**
        - Precision, recall, and F1-score for class 0: 0.75, 0.86, 0.80
        - Precision, recall, and F1-score for class 1: 0.64, 0.47, 0.54
    - **Confusion Matrix Heatmap:** The KNN model's confusion matrix reveals a higher number of false negatives, indicating that the model struggles with correctly identifying class 1. This suggests that KNN may not be the best choice for this dataset without further tuning.



Conclusion

The Random Forest model emerged as the best-performing algorithm in this study, achieving an accuracy of 93.02%. This underscores the power of ensemble methods in handling structured medical data. Decision Trees also demonstrated strong interpretability and competitive accuracy, making them suitable for scenarios requiring explainability. Future research should focus on hyperparameter tuning and exploring advanced algorithms to further improve predictive performance and support early Alzheimer's diagnosis.

Reference

- Dataset Source: Alzheimer's Disease Prediction