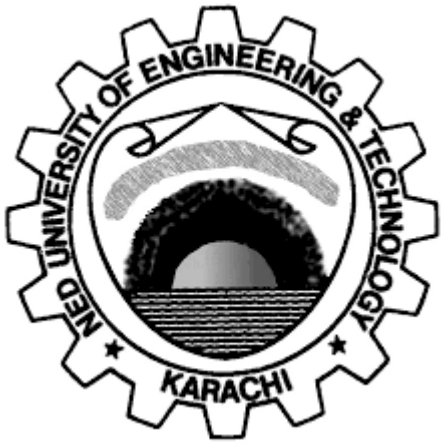
OEL REPORT



GROUP MEMBERS:

NAME	ROLL NO
1. Muhammad Sameed Fayiz	CS-21127
2. Muhammad Ali	CS-21104

Machine Learning CS-324

TE CS BATCH 2021

SPRING SEMESTER 2024

Overview

This project focuses on analyzing and predicting heart attacks using advanced machine learning techniques. The application is built with Streamlit, providing an interactive interface for users to explore data and make heart attack predictions. The project involves comprehensive exploratory data analysis and the implementation of two predictive models: logistic regression and neural networks. Both models are implemented from scratch and using popular Python libraries such as scikit-learn and TensorFlow. The final project was hosted for easy access and demonstration.

Setting Up the Project

Step 1: Create a Python Virtual Environment

To ensure a clean and isolated environment for the project, it is recommended to create a Python virtual environment. This can be done using the following commands:

```
python -m venv env
source env/bin/activate # On Windows use `env\Scripts\activate`
```

Step 2: Install Requirements

Make sure you have all the necessary libraries installed. You can do this by running:

```
pip install -r requirements.txt
```

Step 3: Launch the Application

Start the Streamlit app by running:

```
streamlit run index.py
```

Step 4: Explore and Predict

Dive into the data, train models, and make predictions using the interactive interface.

Dataset Description

The Heart Attack dataset includes the following features related to patient heart health indicators:

- **Age**: Age of the patient
- **Sex**: Gender of the patient (1 = male, 0 = female)
- **cp**: Chest pain type (4 values)
- **trestbps**: Resting blood pressure (in mm Hg)
- **chol**: Serum cholesterol in mg/dl
- **fbs**: Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
- **restecg**: Resting electrocardiographic results (0, 1, 2)
- **thalach**: Maximum heart rate achieved
- **exang**: Exercise-induced angina (1 = yes, 0 = no)
- **oldpeak**: ST depression induced by exercise relative to rest
- **slope**: The slope of the peak exercise ST segment
- **ca**: Number of major vessels (0-3) colored by fluoroscopy
- **thal**: Thalassemia (3 = normal; 6 = fixed defect; 7 = reversible defect)
- **target**: Presence of heart disease (1 = yes, 0 = no)

Exploratory Data Analysis (EDA)

1. Data Summary

We began by loading the dataset and summarizing its structure:

- Shape: The dataset consists of 303 rows and 14 columns.
- Missing Values: There were no missing values in the dataset.
- Descriptive Statistics: We computed the mean, standard deviation, and other descriptive statistics for each feature.

2. Data Visualization

- Histograms: We plotted histograms for each feature to understand their distributions.
- Correlation Heatmap: We created a heatmap to visualize correlations between features and the target variable.
- Pairplot: A pairplot was used to explore relationships between features and the target variable.

3. Feature Relationships

- Age and Heart Attack: We observed that the likelihood of a heart attack increases with age.
- Sex and Heart Attack: Males had a higher prevalence of heart attacks compared to females.
- Chest Pain Type: Certain types of chest pain were more strongly associated with heart attacks.

Data Preprocessing

1. Encoding Categorical Variables

We encoded categorical variables such as `cp`, `restecg`, `slope`, `ca`, and `thal` using one-hot encoding.

2. Feature Scaling

We applied standardization to the dataset to ensure that all features contribute equally to the model's performance.

3. Splitting the Data

The data was split into training and testing sets using an 80-20 split.

Predictive Modeling

Neural Network Models

The neural network consisted of an input layer, two hidden layers with ReLU activation, and an output layer with sigmoid activation.

From Scratch

The neural network model was implemented from scratch to illustrate the basic principles of deep learning. This model achieved an accuracy of 0.868.

TensorFlow

A more advanced neural network model was implemented using TensorFlow. This model achieved an accuracy of 0.855.

Classification Reports:

Neural Network (From Scratch):

	precision	recall	f1-score	support
Class 0	0.8378	0.8857	0.8611	35
Class 1	0.8974	0.8537	0.8750	41
accuracy	-	-	0.8684	76
macro avg	0.8676	0.8697	0.8681	76
weighted avg	0.8700	0.8684	0.8686	76

Neural Network (TensorFlow):

	precision	recall	f1-score	support
Class 0	0.8158	0.8857	0.8493	35
Class 1	0.8974	0.8293	0.8608	41
accuracy	-	-	0.8553	76
macro avg	0.8553	0.8575	0.8550	76
weighted avg	0.8584	0.8553	0.8555	76

Logistic Regression Models

From Scratch

The logistic regression model was implemented from scratch to provide a foundational understanding of this classic machine learning technique. This model achieved an accuracy of 0.882.

Scikit-learn

A logistic regression model was also implemented using the scikit-learn library for a more efficient and robust solution. This model achieved an accuracy of 0.842.

Classification Reports:

Logistic Regression (From Scratch):

	precision	recall	f1-score	support
Class 0	0.8824	0.8571	0.8696	35
Class 1	0.8810	0.9024	0.8916	41
accuracy	-	-	0.8816	76
macro avg	0.8817	0.8798	0.8806	76
weighted avg	0.8816	0.8816	0.8814	76

Logistic Regression (Scikit-learn):

	precision	recall	f1-score	support
Class 0	0.8286	0.8286	0.8286	35
Class 1	0.8537	0.8537	0.8537	41
accuracy	-	-	0.8421	76
macro avg	0.8411	0.8411	0.8411	76
weighted avg	0.8421	0.8421	0.8421	76

Streamlit Application

To make the project interactive and user-friendly, we developed a Streamlit application that allows users to explore the dataset, visualize relationships, and interact with the models. The application was hosted and made accessible for demonstration purposes.

Features of the Streamlit Application

- **Data Exploration**: Users can explore the dataset and view descriptive statistics.

- **Visualization**: Interactive plots and charts to visualize data distributions and relationships.

- **Model Predictions**: Users can input new data and get predictions from the logistic regression and neural network models.

Conclusion

This project successfully demonstrated the process of EDA, data preprocessing, and model implementation using logistic regression and neural networks on the Heart Attack dataset. By implementing models both from scratch and using popular libraries, this project offers a comprehensive learning experience and a practical application for real-world use. The Streamlit application provided an interactive platform for users to engage with the analysis and predictions. Future work could involve optimizing the neural network architecture and exploring other advanced machine learning techniques to further improve prediction accuracy.