Heart Disease Prediction

Satvik Kundargi

Introduction

- Heart disease is one of the leading causes of death worldwide.
- Predicting heart disease based on clinical features can help healthcare professionals take preventive measures.
- This project focuses on building machine learning models to predict whether a person has heart disease based on medical data.
- **Goal:** Develop a predictive model using patient data to identify heart disease with high accuracy.
- Tools used: Python, machine learning libraries (Scikit-learn, Keras), Gradio (for the web interface).
- Overview of steps: Data cleaning, model building, evaluation, and deployment through an interactive web application.

Problem Definition

- **Problem**: Early detection of heart disease to prevent severe outcomes.
- **Importance**: Late diagnosis can lead to complications, increased healthcare costs, and fatalities.
- **Business Impact**: A reliable predictive model could enhance diagnostic accuracy and guide early interventions.
- **Solution**: Use machine learning to build a model that predicts heart disease based on clinical data.

Dataset Overview

- **Dataset**: Kaggle Heart Disease Dataset
- **Source**: Available from kaggle Repository.
- **Age:** The patient’s age in years.
- **Sex:** The gender of the patient (1 = Male, 0 = Female).
- **CP (Chest Pain Type):** Type of chest pain experienced by the patient (0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic).
- **Trestbps (Resting Blood Pressure):** Resting blood pressure (in mm Hg) when the patient was admitted to the hospital.
- **Chol (Serum Cholesterol):** The patient’s cholesterol level in mg/dL.
- **FBS (Fasting Blood Sugar):** Indicates whether the patient’s fasting blood sugar is above 120 mg/dL (1 = true, 0 = false).

Dataset Overview

- **Restecg (Resting Electrocardiographic Results):** Results of the resting electrocardiogram (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy).
- **Thalach (Maximum Heart Rate Achieved):** The highest heart rate achieved during exercise stress testing.
- **Exang (Exercise Induced Angina):** Whether the patient experienced angina (chest pain) induced by exercise (1 = yes, 0 = no).
- **Oldpeak:** Depression in ST segment relative to rest (measured in mm), reflecting exercise-induced ischemia.
- **Slope (Slope of ST Segment):** The slope of the peak exercise ST segment (0 = upsloping, 1 = flat, 2 = downsloping).
- **Ca (Number of Major Vessels Colored by Fluoroscopy):** Number of major blood vessels (0–3) colored by a fluoroscopy procedure.
- **Thal:** Results of a thallium stress test (1 = normal, 2 = fixed defect, 3 = reversible defect).
- **Target:** The output variable indicating the presence of heart disease (1 = disease, 0 = no disease).

Exploratory Data Analysis (EDA)

- **Objective of EDA:** To understand the dataset’s structure, distribution, and relationships between variables, and to gain insights before modeling.
- **Key findings:**
- **Distribution of Age:** Most patients are in their 50s and 60s.
- **Sex:** The majority of patients are male, which reflects higher heart disease rates in men.
- **Cholesterol levels:** Wide variation across patients, with outliers identified.
- **Correlation analysis:** Strong correlations between features like Chest Pain Type and heart disease status.

Feature Engineering

- **Age group**
- **Higher heart rate**
- **Interaction**
- **Handling Categorical Variables**
- **Normalization/Scaling of Numerical Variables**
- **Binarization of Fasting Blood Sugar (FBS)**

Machine Learning Models Tried

- **Logistic Regression**:
  - Reason: Simple and interpretable model for binary classification.
- **Naive Bayes**:
  - Reason: Fast, performs well with small datasets.
- **Neural Network (MLP using Keras)**:
  - Reason: Complex model that can capture non-linear relationships.

Model Evaluation and Metrics

- **Metrics Used**:
- **Accuracy**: Overall correctness of the model.
- **Precision**: The model’s ability to identify true positives (important for healthcare).
- **Recall**: Ability to detect actual cases of heart disease (minimizing false negatives is critical).
- **F1-Score**: A balance between precision and recall.
- **Performance**:
- Logistic Regression: 78% accuracy, Precision: 84%
- Naive Bayes: 75% accuracy, Precision: 79%
- Neural Network: 95% accuracy, Precision: 97%
- **Trade-offs**: While the Neural Network performs better in accuracy and precision, Logistic Regression provides better interpretability.

Web Interface for Heart Disease Prediction

- The web interface was built using **Gradio** to allow users (doctors, patients) to interact with the model.
- **User-friendly inputs:** Users can input data like age, cholesterol levels, and chest pain type using sliders and dropdowns.
- **Prediction Results:** Once the user submits their data, the model predicts whether the individual is at risk for heart disease.
- **Purpose:** This interface can be deployed in healthcare settings to assist with patient diagnosis.

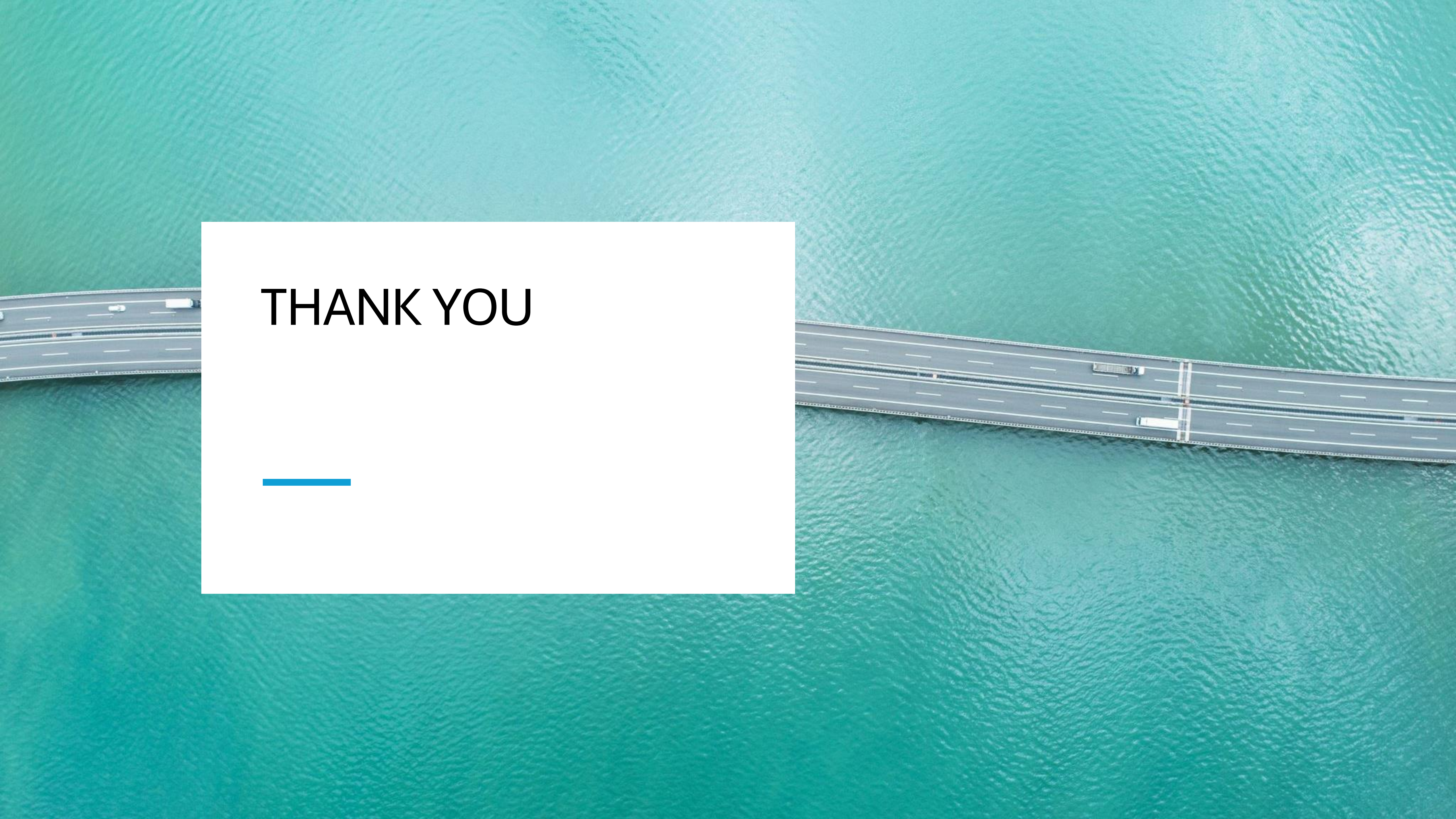
FUTURE WORK

- **Potential improvements:**
- Incorporating additional data, such as lifestyle factors (diet, exercise) or genetic information.
- Adding Explainable AI (XAI) techniques to provide insights into the model’s predictions, which could enhance trust in healthcare applications.
- Testing with a larger and more diverse dataset to improve model generalization.
- **Exploring more complex architectures** for the neural network, such as deeper layers or different activation functions.
- Collaborating with domain experts to ensure the model’s clinical relevance and applicability.

Conclusion

- **Summary:** The project successfully developed a predictive model for heart disease using clinical data.
- **Best Model:** The MLP Neural Network outperformed others, making it the best option for the given task.
- **Real-world impact:** The model, combined with the Gradio web interface, could assist healthcare professionals in making quicker and more accurate diagnoses.


Any Queries ?



THANK YOU