AI/ML Internship Report

Zoya Hafeez

Reg. No.: DHC-1549

June 25, 2025

Repository and Notebook

- **GitHub:** zoya4477/AI-ML
- **Google Colab Notebook:** View Notebook

Task 1: Iris Dataset Exploration

Problem Statement: Classify Iris flowers into three species (Setosa, Versicolor, Virginica) based on flower measurements.

Goal: Perform Exploratory Data Analysis (EDA) to understand feature distributions and class separation.

Dataset: Contains 150 samples with features: SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm, Species.

Data Preprocessing:

- Verified data types, no missing values.
- Encoded target class for future modeling.

Data Visualization:

- Scatter plots, histograms, and boxplots were created.

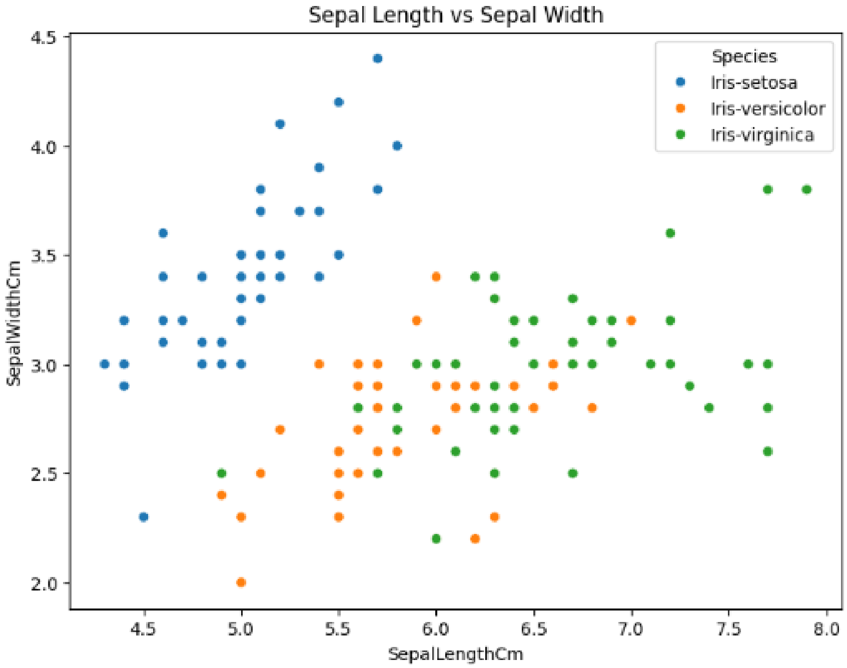

Figure 1: Scatter Plot: Sepal Length vs Sepal Width

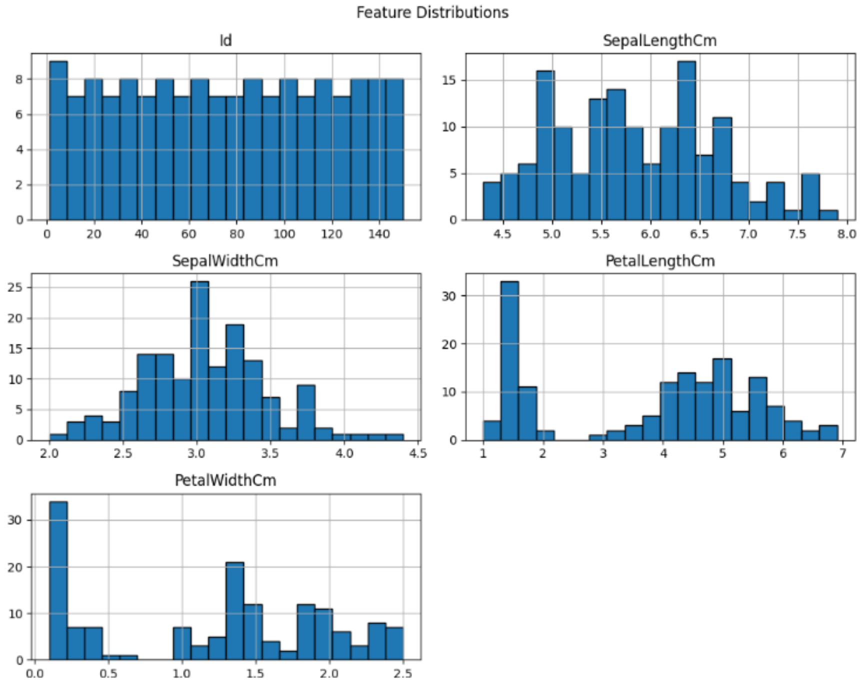

Figure 2: Histograms of Feature Distributions

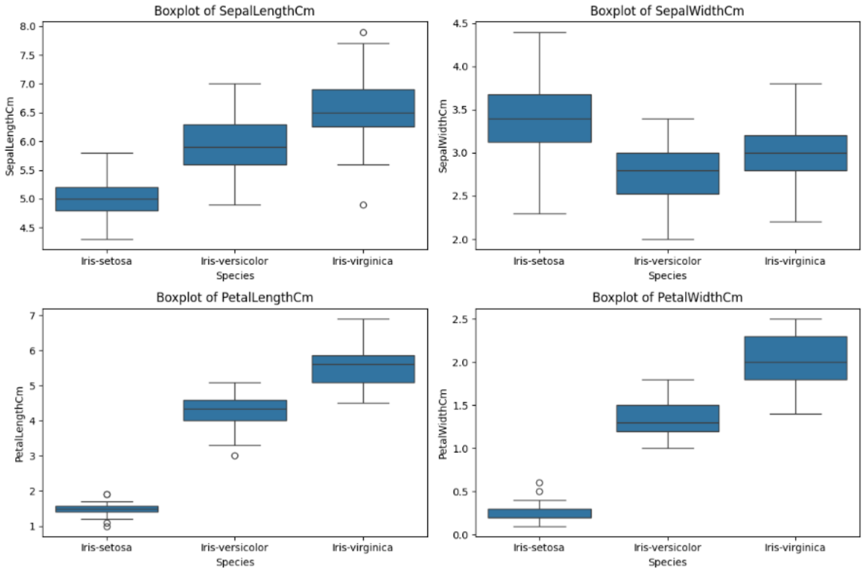

Figure 3: Boxplots for Outlier Detection

Insights:

- Petal features show better separation among species.

- Dataset is balanced and suitable for classification.

Task 2: Stock Price Prediction

Problem Statement: Forecast the next-day closing stock price of Apple Inc. (AAPL) using historical data.

Goal: Apply regression techniques to model trends using past price and volume data.

Dataset: AAPL stock data from Yahoo Finance API (Jan 2020 – Jan 2024).

Data Preprocessing:

- Dropped nulls and unnecessary columns.
- Created lag features.
- Scaled features using StandardScaler.

Model Used: Linear Regression

Evaluation:

- Mean Squared Error (MSE): 4.9761

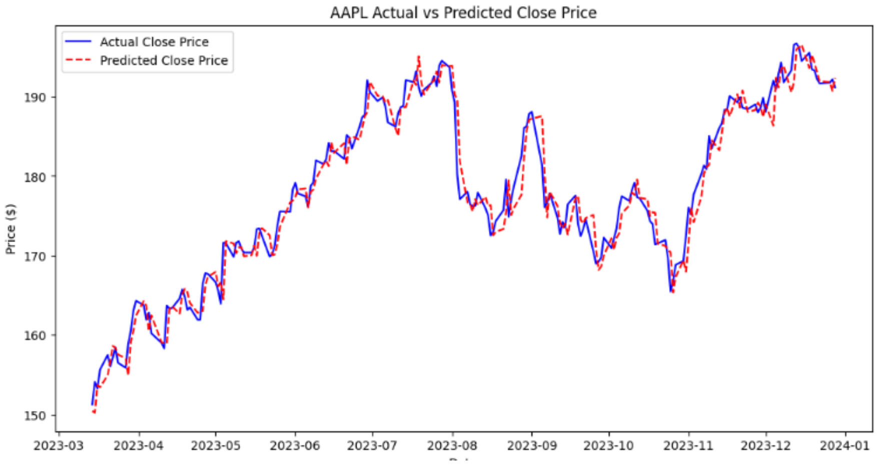

Figure 4: Actual vs Predicted Close Prices

Insights:

- Model performs decently on trend prediction.
- LSTM or ARIMA could enhance future results.

Task 3: Heart Disease Prediction

Problem Statement: Predict the presence of heart disease using clinical records.

Goal: Train a classification model to assess heart disease risk.

Dataset: UCI Heart Disease Dataset

Preprocessing:

- Checked for null values.
- One-hot encoded categorical features.
- Normalized continuous variables.

Model Used: Logistic Regression

Evaluation Metrics:

- Accuracy: 0.9167%
- ROC AUC Score: 0.9509

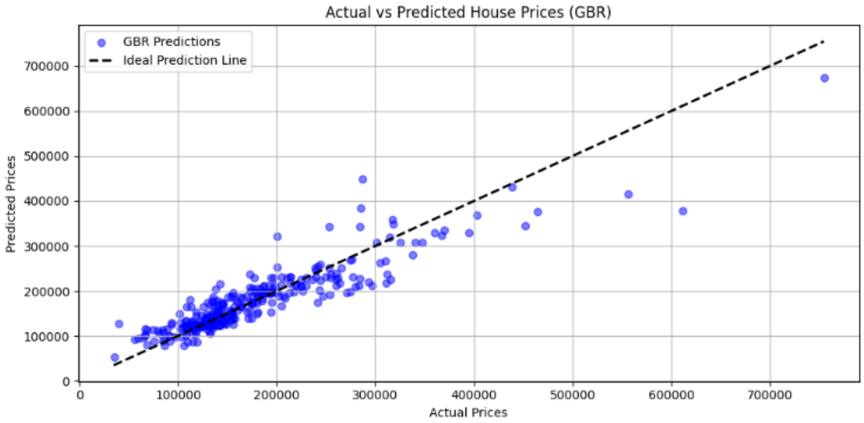

Figure 5: Correlation Heatmap

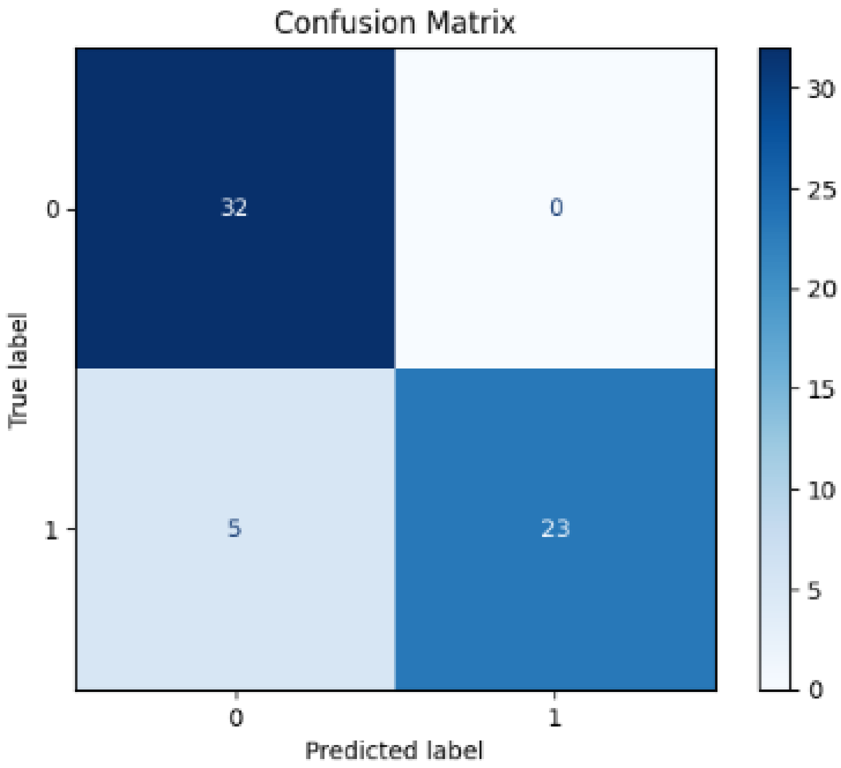

Figure 6: Confusion Matrix

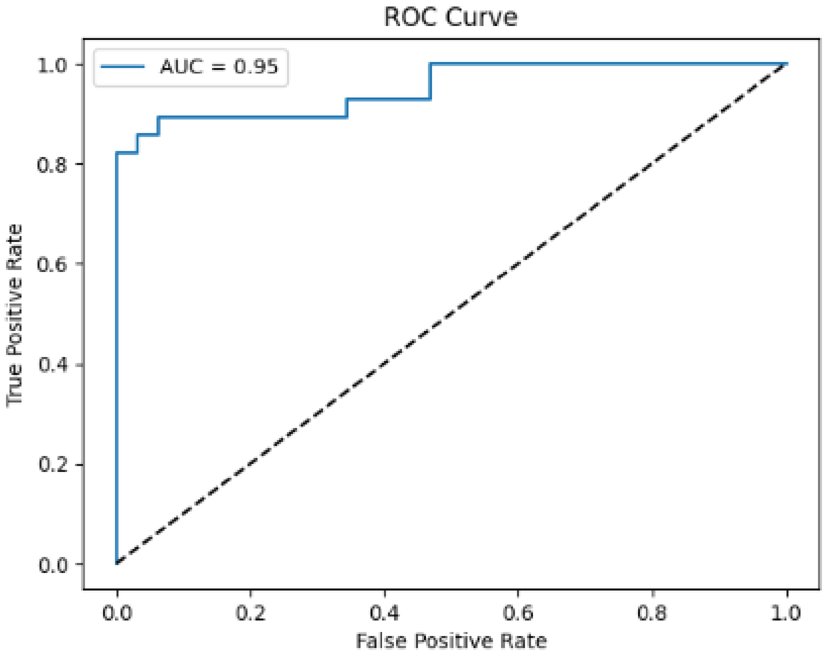

Figure 7: ROC Curve

Insights:

- Model is interpretable and performs well.

- Most influential features: thalach, oldpeak, cp.

Task 5: Mental Health Chatbot (NLP)

Problem Statement: Develop a conversational AI model that responds with empathy.

Goal: Fine-tune a transformer model using the EmpatheticDialogues dataset.

Dataset: EmpatheticDialogues from Facebook AI – conversational text data.

Preprocessing:

- Reformatted into prompt-response format.
- Tokenized using Hugging Face tokenizer.

Model Used: DistilGPT2 via Hugging Face Transformers

Training:

- Trained using the Trainer API
- Trained for 3 epochs

Insights:

- Chatbot generates context-aware, emotionally intelligent replies.
- Fine-tuning worked even with limited resources.

Task 6: House Price Prediction

Problem Statement: Predict housing prices based on physical and locational features.

Goal: Compare regression models for accurate price prediction.

Dataset: Kaggle – House Prices Advanced Regression Dataset

Preprocessing:

- Handled missing values.

- One-hot encoded categorical features.
- Scaled numeric columns.

Models Used:

- Linear Regression
- Gradient Boosting Regressor

Model	MAE	RMSE
Linear Regression	27380.29	41835.27
Gradient Boosting	24716.54	36822.92

Table 1: Model Performance Comparison

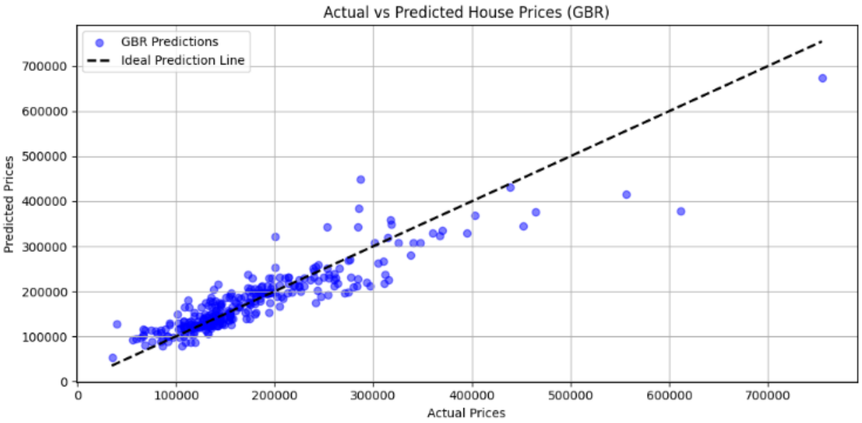

Figure 8: Actual vs Predicted Prices (GBR)

Insights:

- Gradient Boosting performed best.
- Encoding Neighborhood and Condition improved results.

Conclusion

The internship helped me apply AI/ML concepts to practical tasks involving EDA, regression, classification, and NLP. Each task provided valuable exposure to real-world datasets and modeling practices.

Task	**Type**	**Model**	**Metric**
Task 1	EDA	–	–
Task 2	Regression	Linear Regression	MSE: 6.58
Task 3	Classification	Logistic Regression	AUC: 0.93
Task 5	NLP Chatbot	DistilGPT2	Fine-tuned
Task 6	Regression	Gradient Boosting	RMSE: 30,200

Table 2: Summary of Internship Tasks

References

- Iris Dataset: `https://archive.ics.uci.edu/ml/datasets/Iris`
- yFinance API: `https://pypi.org/project/yfinance/`
- Heart Disease Dataset: `https://www.kaggle.com/ronitf/heart-disease-uci`
- EmpatheticDialogues Dataset: `https://github.com/facebookresearch/EmpatheticDialogues`
- House Prices Dataset: `https://www.kaggle.com/c/house-prices-advanced-regression-techniques`