Appendix A

Python Code Used for Data Preparation and Logistic

Regression Modeling

This project utilized a combination of analytical tools to support data preparation, modeling, and visualization.

A1. Importing Libraries and Preparing the Environment

```python
import pandas as pd
import numpy as np
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_auc_score
```

A2. Loading the Cleaned Dataset

```python
df = pd.read_csv('/mnt/data/diabetic_data_na_removed.csv')
```

A3. Recoding the Readmission Variable

A binary variable was created to represent 30-day readmission.

Appendix B

Notes on Analytical Tools Used

This project utilized a combination of analytical tools to support data preparation, modeling, and visualization.

B1. R

R was used to perform the initial data cleaning by replacing all question marks that denoted NA with actual NA cells so other software, such as Orange, would correctly import the data.

B2. Python

Python (pandas, NumPy, scikit-learn) was used for:

- Data cleaning

- Feature engineering

- Logistic regression modeling

- Extracting model coefficients

- Computing odds ratios

- Calculating the ROC-AUC score

Python provided statistical rigor and flexibility in manipulating the dataset.

B3. Orange Data Mining

Orange was used for:

- Building the visual workflow diagram

- Running machine-learning models (e.g., random forest, gradient boosting)

- Generating ROC curves for classifier comparison

- Visualizing the overall modeling pipeline

- Data Cleaning

This tool allowed us to compare model families and clearly illustrate the modeling steps and allowed us to have a direct comparison between all 3 models tested.

B4. Visualization Tools

Built-in plotting tools in Orange were used to generate:

- Age distribution

- Gender distribution

- Race distribution

- Number of lab procedures

- Number of medications

- Length of stay

- Weight distribution

- Medication change indicators

These visualizations support the exploratory data analysis presented in the report.

Appendix C

Figures and Screenshots

1. Age Distribution Histogram

2. Gender Distribution Bar Chart

3. Race Distribution Bar Chart

4. Number of Lab Procedures Histogram

5. Number of Medications Histogram

6. Time in Hospital Histogram

7. Weight Distribution Histogram

8. Medication Change Distribution Chart

9. Orange Workflow Diagram

10. ROC Curve – Target 0

11. ROC Curve – Target 1