**D603 - Machine Learning**

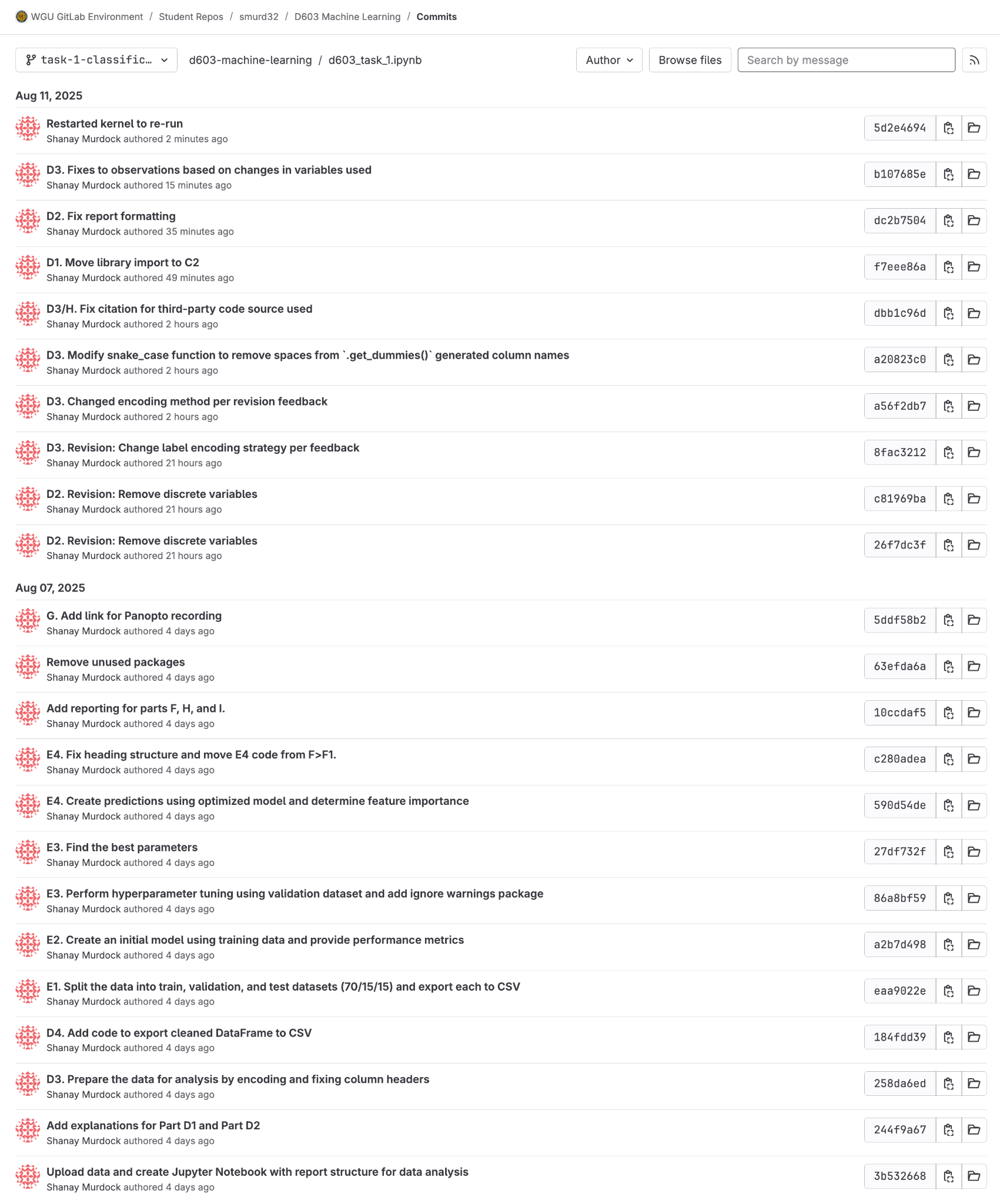
**RFN1 Task 1: Classification Data Mining Models**

Prepared by Shanay Murdock

## A. GitLab Repository

<https://gitlab.com/wgu-gitlab-environment/student-repos/smurd32/d603-machine-learning/-/tree/task-1-classification?ref_type=heads>

Screenshot of commit history:



## B. Purpose of the Data Mining Report

### B1. Proposal of Question

**Business Question:** Can we predict which patients are most likely to be readmitted within 30 days of discharge so that our hospital chain can implement targeted intervention programs to improve patient outcomes and reduce readmission penalties?

### B2. Defined Goal

**Goal:** Develop a Random Forest Classification model that identifies the most important factors influencing the prediction of patient readmission risk, enabling early identification of high-risk patients for targeted intervention programs. By identifying these factors, we can provide targeted intervention initiatives that can improve the patient experience and reduce the financial penalties from the Centers for Medicare and Medicaid Services (CMS). Freeing up that funding can go toward additional hospital resources or medical care initiatives.

## C. Justification of Classification Method

### C1. Explanation of Classification Method

Random Forest Classification works by creating multiple decision trees that each make independent predictions about patient readmission risk. Each tree is trained on a random sample of the data and uses a random subset of available features, such as age, medical conditions, and hospital stay details. The final prediction comes from combining all the individual tree votes--if most trees predict "readmitted," the patient is classified as high-risk for readmission. This ensemble approach reduces errors and provides more reliable predictions than any single decision tree.

Random Forest Classification is the ideal classification method for this medical dataset because it naturally handles the mix of categorical data (like gender and medical conditions) and continuous data (like age and income) without requiring complicated preprocessing. Most importantly for healthcare applications, it provides clear feature importance rankings that show which patient factors most strongly predict readmission risk. This interpretability allows hospital staff to understand why certain patients are high-risk and develop targeted intervention programs.

The expected outcomes include accurate binary predictions (readmitted versus not readmitted), probability scores for risk assessment, and actionable insights about which patient characteristics to monitor most closely.

### C2. Packages or Libraries List

The follow packages and libraries were chosen to complete the Random Forest Classification:

* pandas: Allows data manipulation and analysis.
* numpy: Provides numerical computations.
* re: RegEx package allows easy renaming of variables for consistency.
* RandomForestClassifier from Scikit-Learn: Acts as the main classification algorithm.
* train\_test\_split from Scikit-Learn: Provides data splits for training and testing to prevent overfitting.
* GridSearchCV from Scikit-Learn: Finds optimal hyperparameters automatically using cross-validation.  
  StratifiedKFold from Scikit-Learn: A cross-validation technique used to split a dataset into training and testing sets, specifically designed to handle similar proportions of samples from each class as the original dataset.
* cross\_val\_score from Scikit-Learn: Provides a more robust performance estimation than single train/test split.
* accuracy\_score from Scikit-Learn: Provides overall correctness but can be misleading if classes are imbalanced.
* precision\_score from Scikit-Learn: Reduces false alarms, important for resource allocation.
* recall\_score from Scikit-Learn: Measures how good the model is at finding all positive cases that actually exist.
* f1\_score from Scikit-Learn: Balances precision and recall.
* roc\_auc\_score from Scikit-Learn: Determines how well the model distinguishes between readmitted and non-readmitted patients.
* confusion\_matrix from Scikit-Learn: Provides detailed breakdown of prediction errors.
* classification\_report from Scikit-Learn: Single function that provides all performance metrics.
* matplotlib: Visualizations
* seaborn: Statistical visualizations



## D. Data Preparation

### D1. Data Preprocessing

**Data Preprocessing Goal:** Convert categorical variables to numerical format to create compatibility with Random Forest algorithm requirements.

Random Forest Classification requires all input features to be in numerical format, but the medical\_clean.csv dataset contains several categorical variables such as Gender (Male/Female/Nonbinary), Marital Status (Divorced/Married/Widowed), and several binary medical conditions (Yes/No) for high blood pressure, stroke, diabetes, etc.

To address this requirement, I use mapping and pandas’ get\_dummies() to transform these text-based categorical variables into numeric representations and Boolean values. For example, I will use mapping to replace the 'Yes' and 'No' to '1' and '0' for any binary categories, and get\_dummies() will be used to encode nominal categories into Boolean values, which can be accepted implicitly by Random Forest Classification as numeric values.

This preprocessing step is essential because Random Forest's underlying decision tree algorithm can only process numerical data when making splits and calculating feature importance scores. By encoding categorical variables appropriately, it is ensured that the Random Forest model can effectively utilize all available patient information to make accurate readmission predictions while maintaining the interpretability of results.

### D2. Dataset Variables

Below are the initial dataset variables retained for the Random Forest Classification analysis to predict patient readmission status:

#### Target Variable:

* ReAdmis: Patient readmitted within 30 days (Yes/No); Categorical (Binary)

#### Predictor Variables:

**Patient Demographics:**

* Age: Patient age in years; Continuous
* Gender: Patient gender (Male/Female/Nonbinary); Categorial
* Income: Annual income of patient; Continuous
* Marital: Marital status of patient (Single/Married/Divorced/Widowed/Never Married); Categorical
* Area: Area type (Rural/Suburban/Urban); Categorical

**Clinical Measurements:**

* VitD\_levels: Vitamin D levels (ng/mL); Continuous
* Initial\_days: Length of initial hospital stays (days); Continuous

**Medical Conditions (Binary Yes/No):**

* HighBlood: High blood pressure diagnosis; Categorical (Binary)
* Stroke: Previous stroke history; Categorical (Binary)
* Overweight: Patient considered overweight; Categorical (Binary)
* Arthritis: Arthritis diagnosis; Categorical (Binary)
* Diabetes: Diabetes diagnosis; Categorical (Binary)
* BackPain: Chronic back pain diagnosis; Categorical (Binary)
* Anxiety: Anxiety disorder diagnosis; Categorical (Binary)
* Asthma: Asthma diagnosis; Categorical (Binary)

**Operational Factors:**

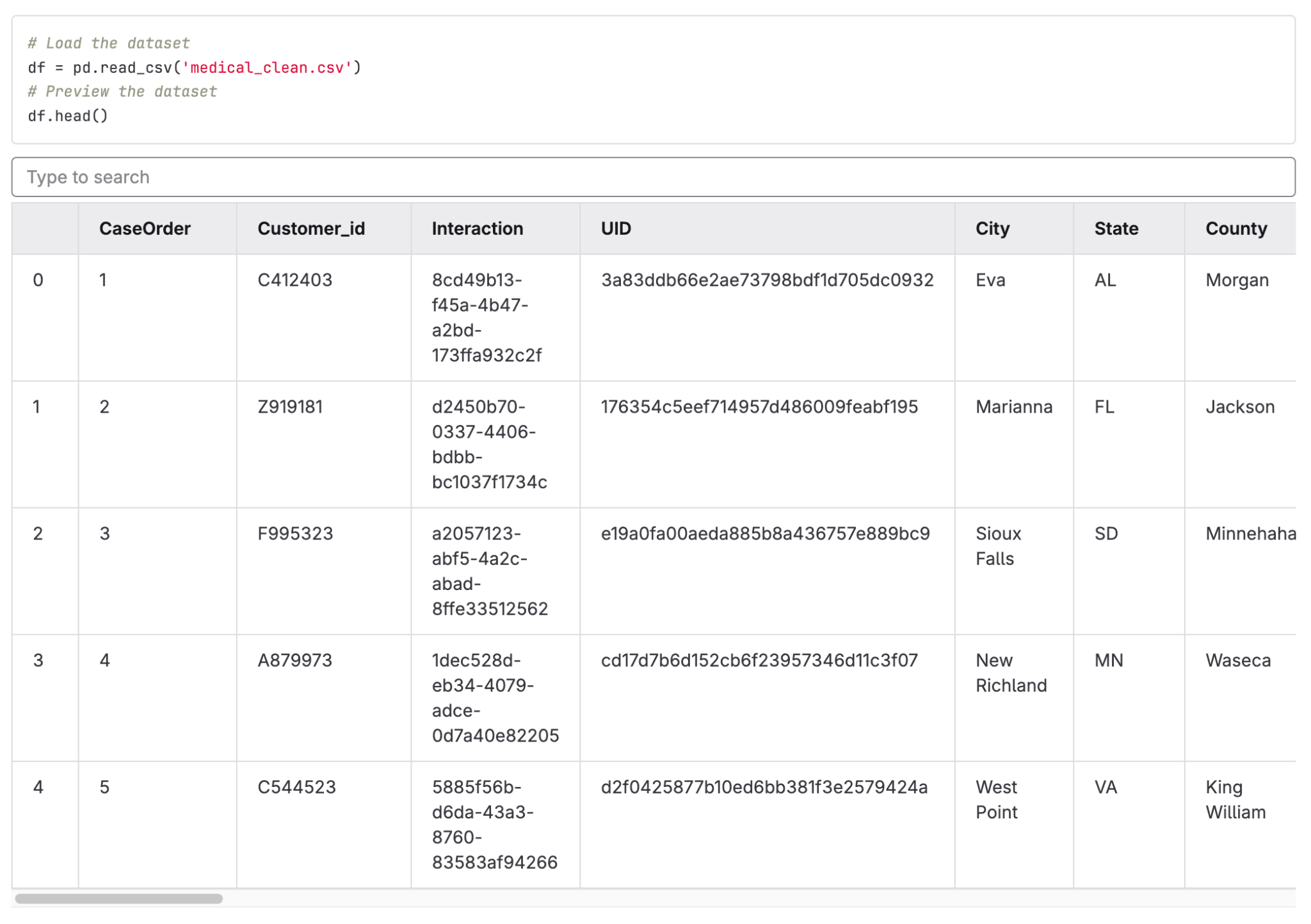
* Initial\_admin: Type of admission (Emergency/Elective/Observation); Categorical
* Services: Primary service received (Blood work/IV/CT scan/MRI); Categorical
* Complication\_risk: Patient complication risk level (High/Medium/Low); Categorical (Ordinal)
* Soft\_drink: Habitually drinks 3+ sodas daily (Yes/No); Categorical (Binary)

**Financial Variables**

* TotalCharge: Average daily charge amount; Continuous
* Additional\_charges: Average additional charges; Continuous

Summary of feature variables: 15 categorical variables; 6 continuous variables

### D3. Steps for Analysis



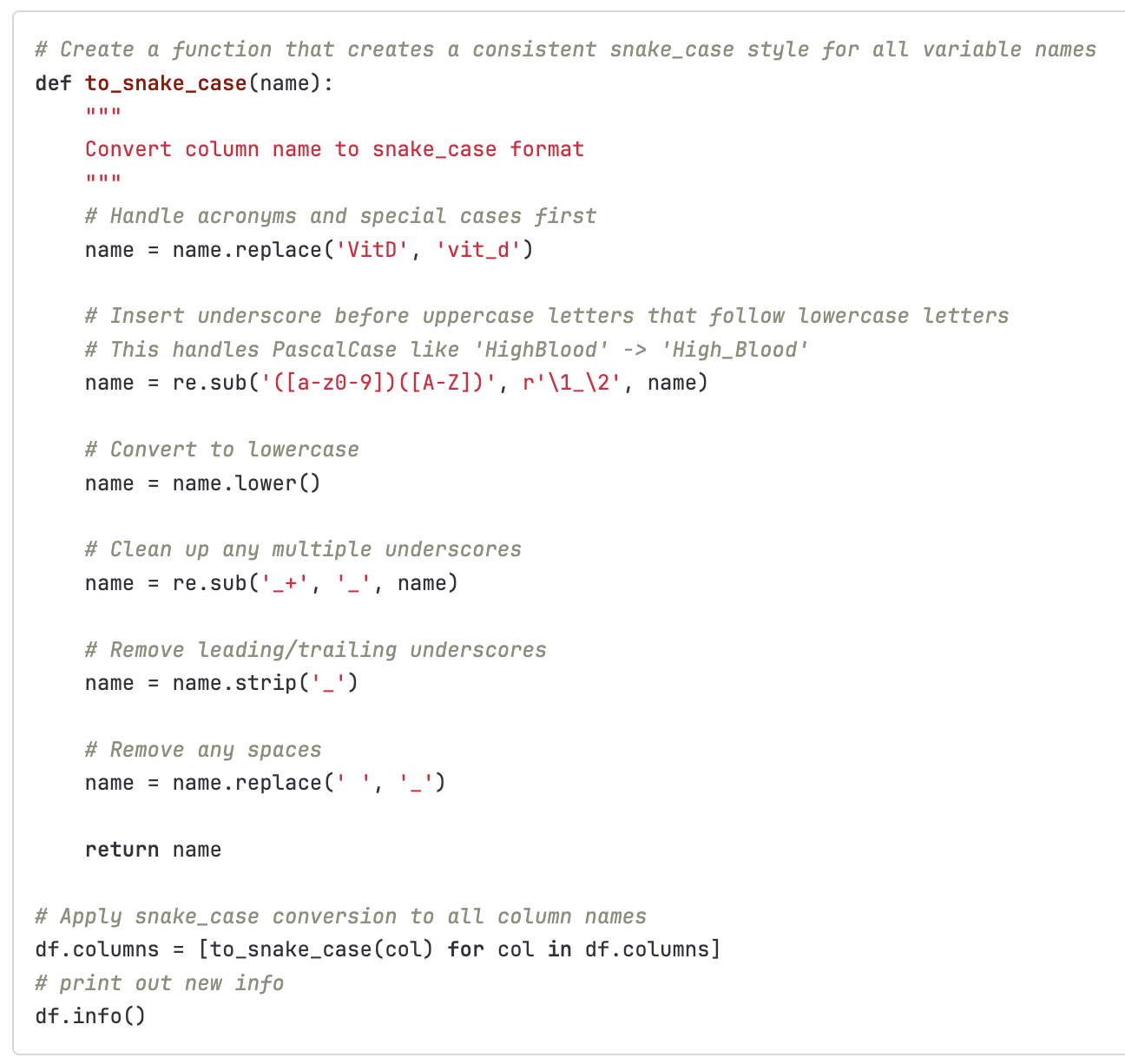
To prepare the data for the Random Forest classification model, I used the .read\_csv() function from pandas to load the CSV and convert it into a DataFrame, making it easier to analyze. The .head() function loads the first five observations of the DataFrame to verify it loaded correctly and provide a full preview of the data.



Out of the original 50 features, I chose a target variable ('ReAdmis') and 21 potential predictive variables that seemed to have the most to do with what may predict if a patient is readmitted to the hospital after discharge.

**Initial Observations:**

* There are no missing values in any of the columns.
* There is a mix of object, int64, and float64 data types.
* There is one target variable and 21 feature variables out of the initial 50 variables.

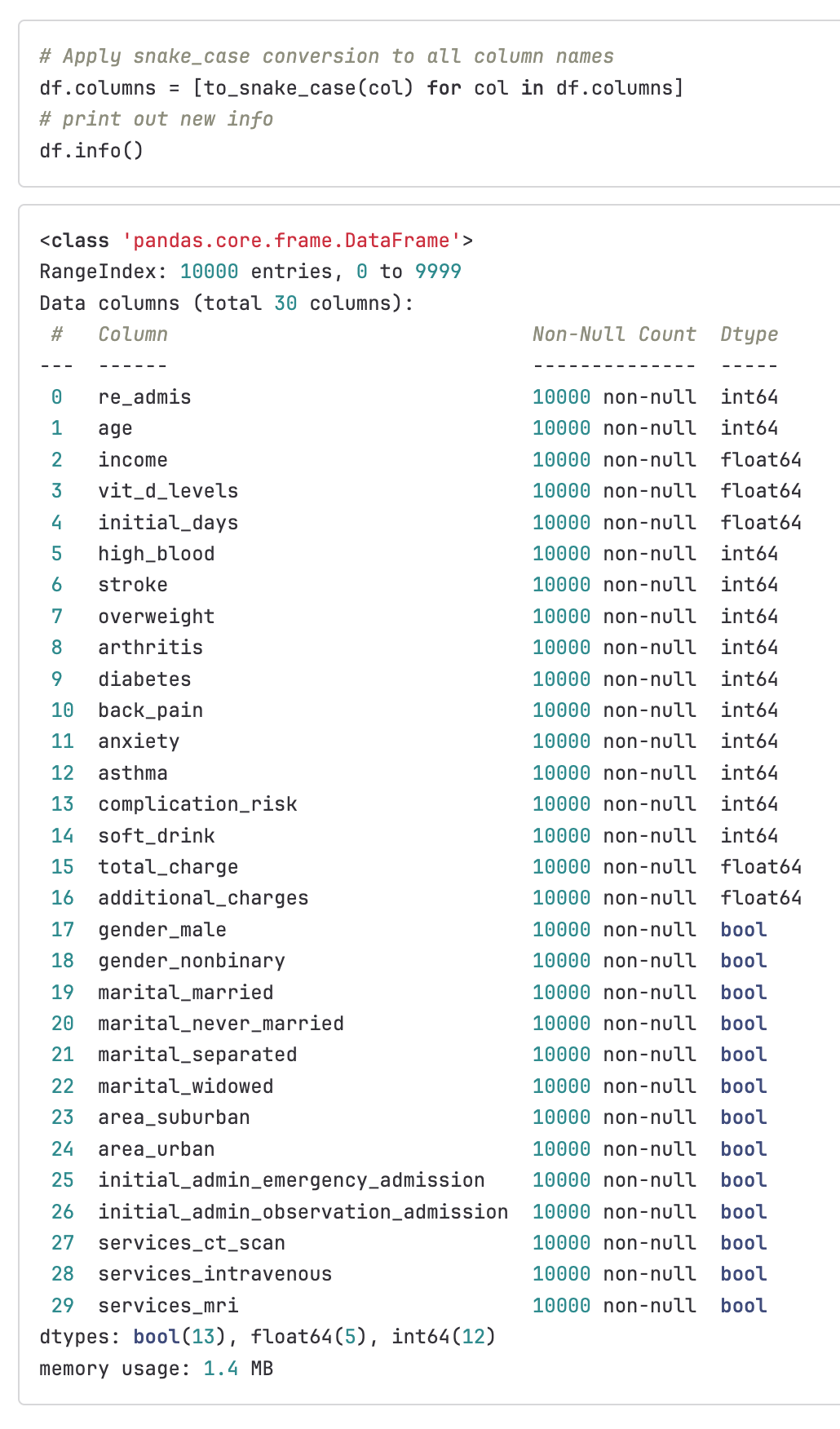


I found the inconsistency of the column names (despite being a "cleaned" dataset) to be difficult to work with, so I created a function that uses regular expressions to fix the column names and convert them all from a mix of snake case, Pascal case, and other miscellaneous cases to a consistent use of snake\_case.



I separated out the binary, ordinal, and nominal categorical variables and encoded them with a mix of mapping and pd.get\_dummies().

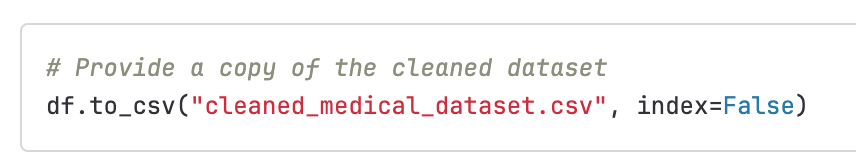
Using DataFrame mapping, I was able to quickly convert binary "Yes" values to 1 and "No" values to 0 and create an order for the complication\_risk values. I used pandas' pd.get\_dummies() to one-hot encode the nominal variables in a way that does not assign meaning to the order and this generated new features. I left the Boolean values as bool types as Random Forest Classification can take bool values without further preprocessing into explicit numeric values.



**Observations:**

* There are no longer any "object" types. All data has been converted to int64, float64, or Boolean as appropriate.

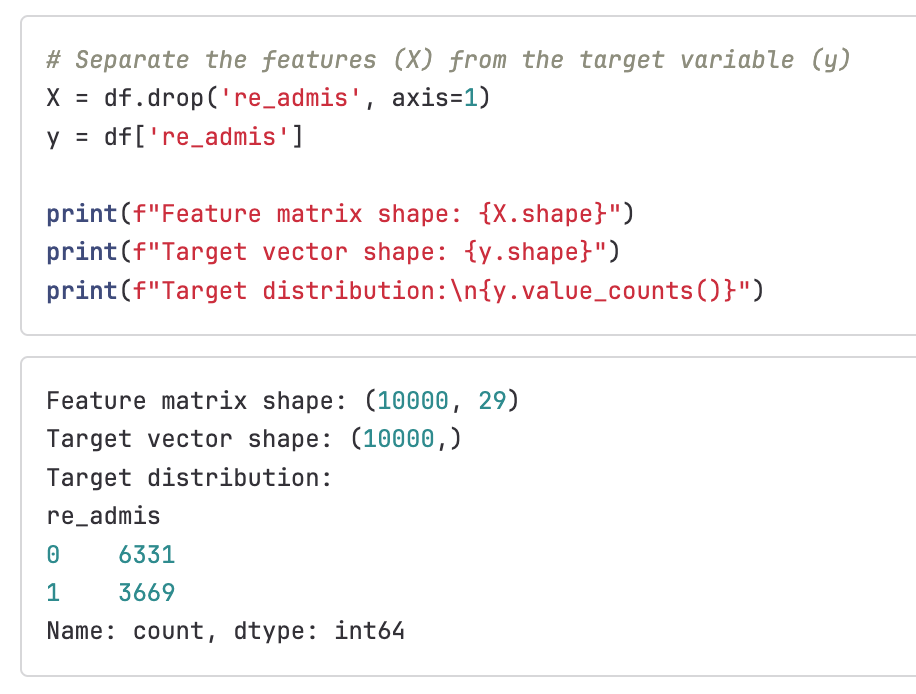
### D4. Cleaned Dataset

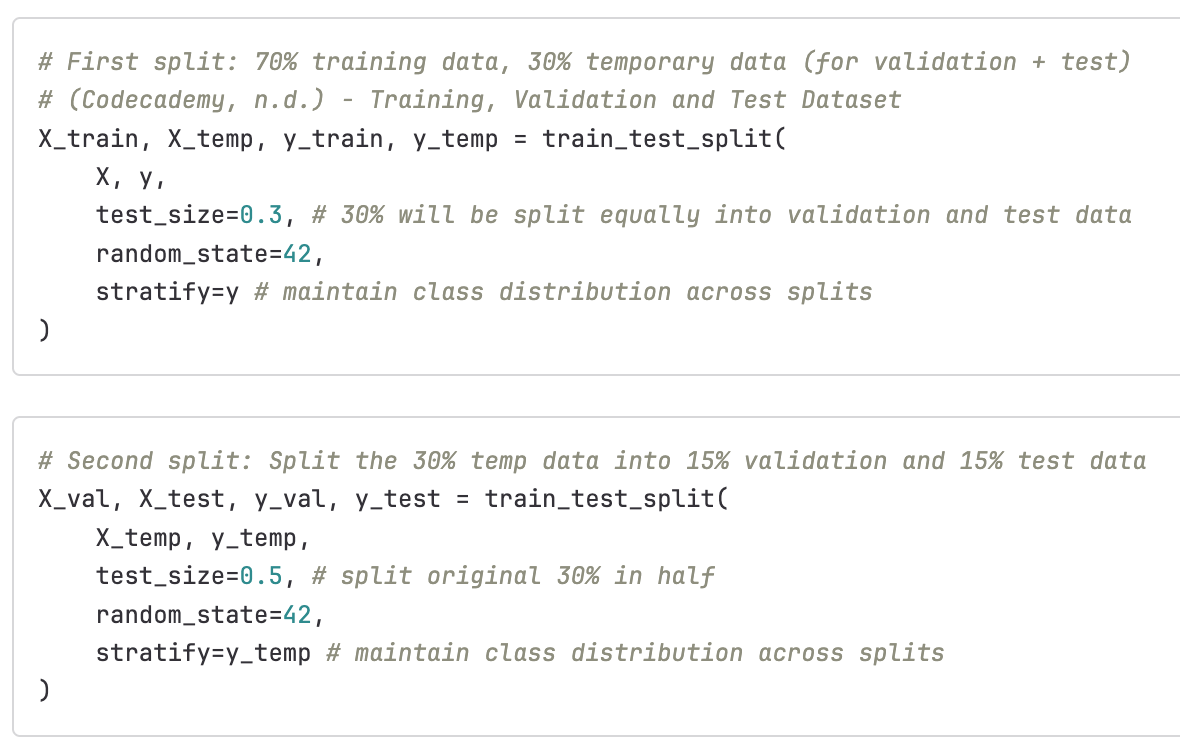


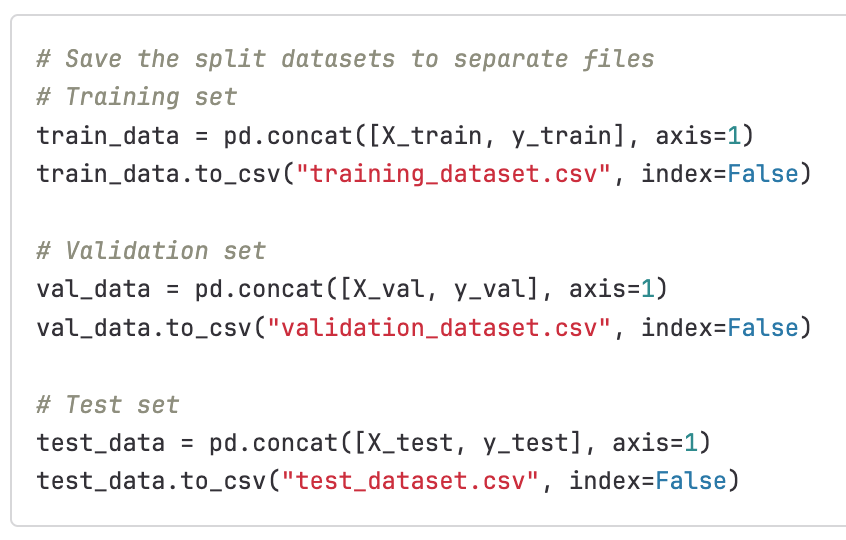
The cleaned dataset is in a separate file named cleaned\_medical\_dataset.csv.

## E. Perform Data Analysis

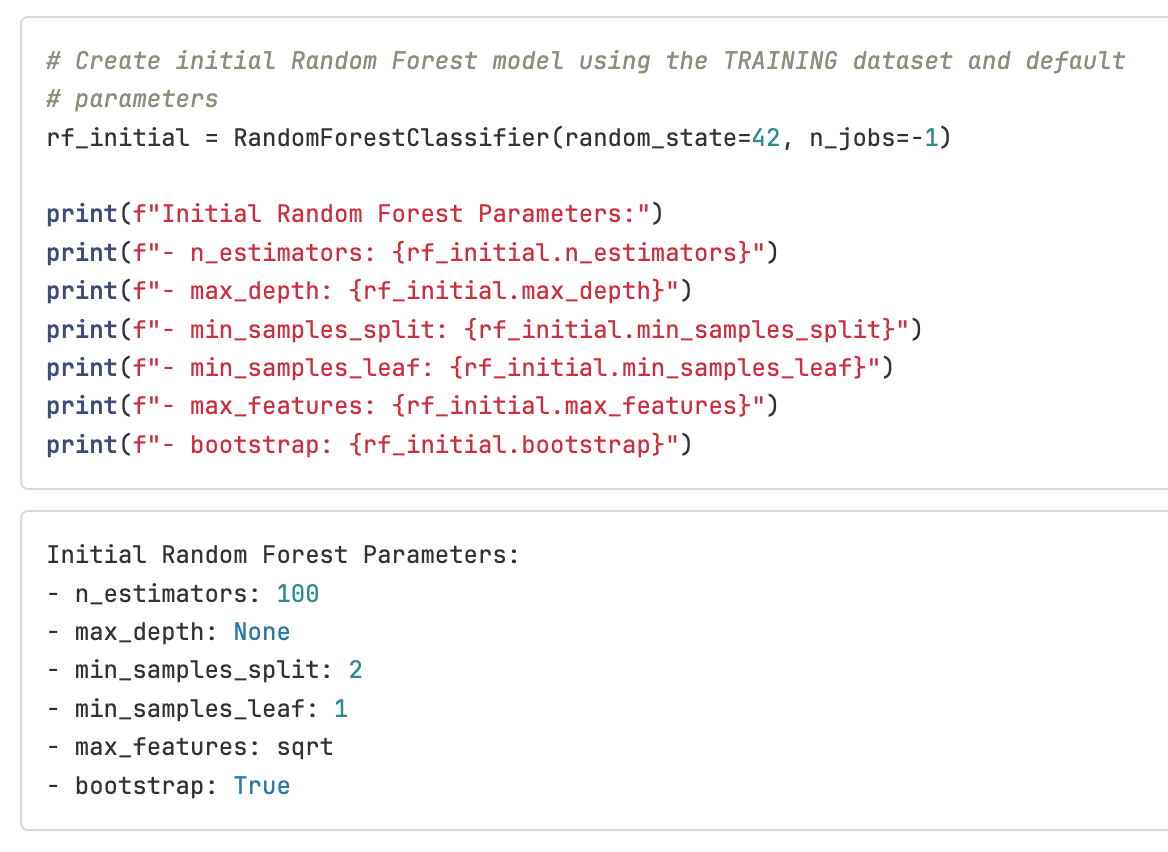
### E1. Splitting the Data

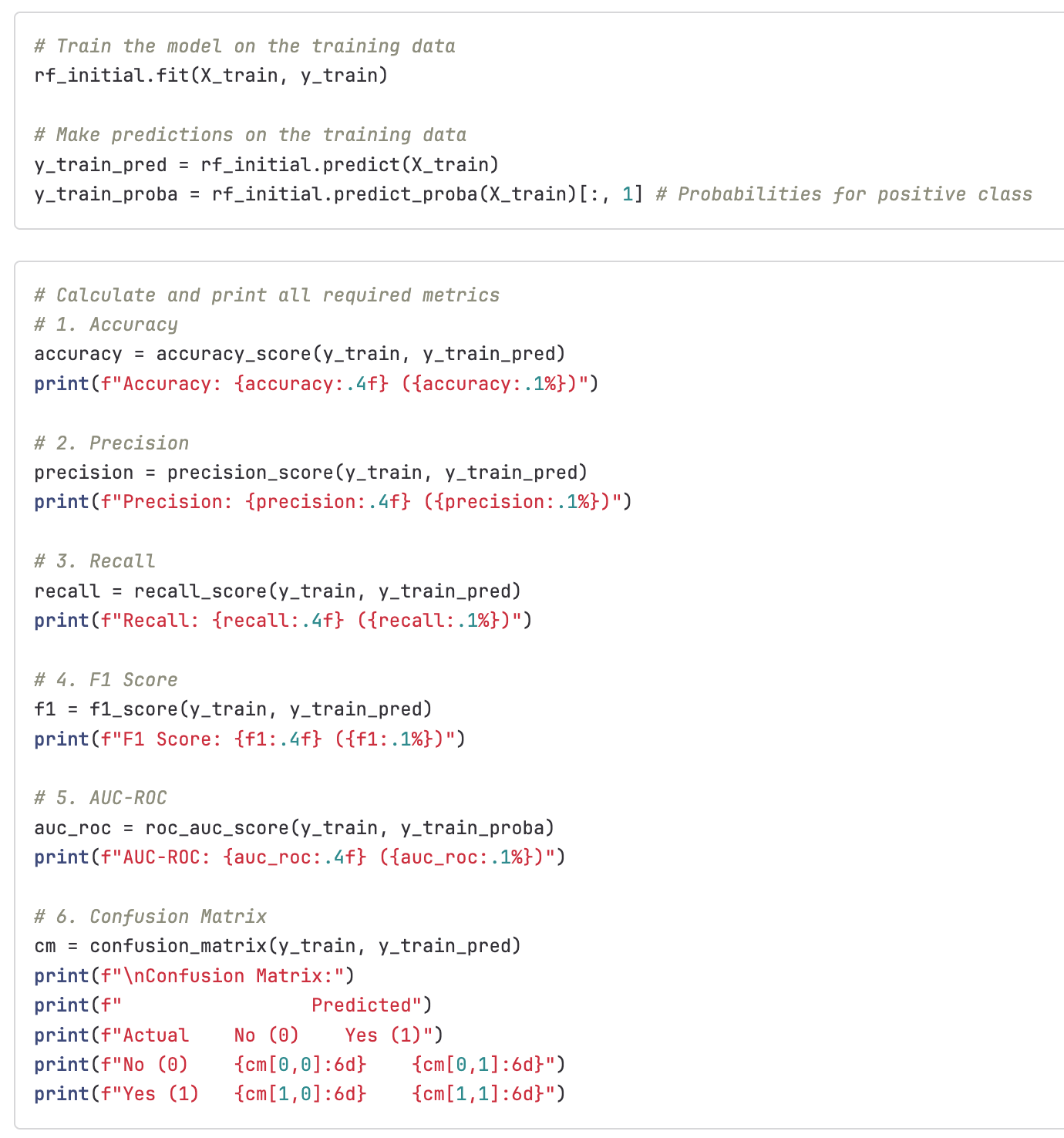




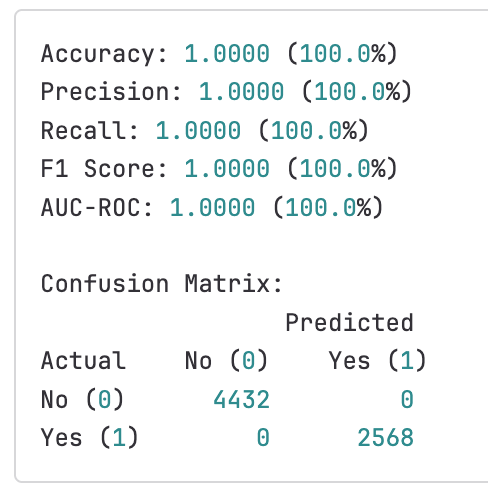


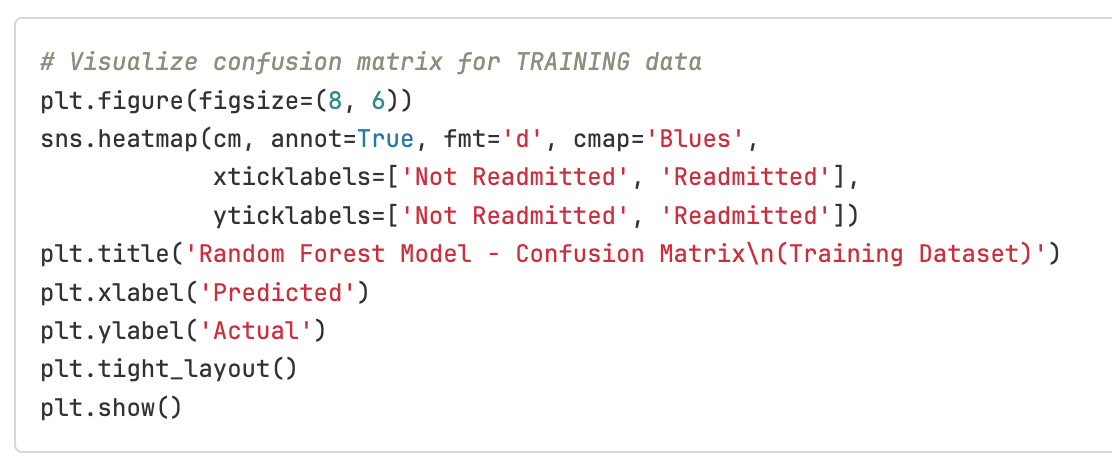
### E2. Initial Model Creation

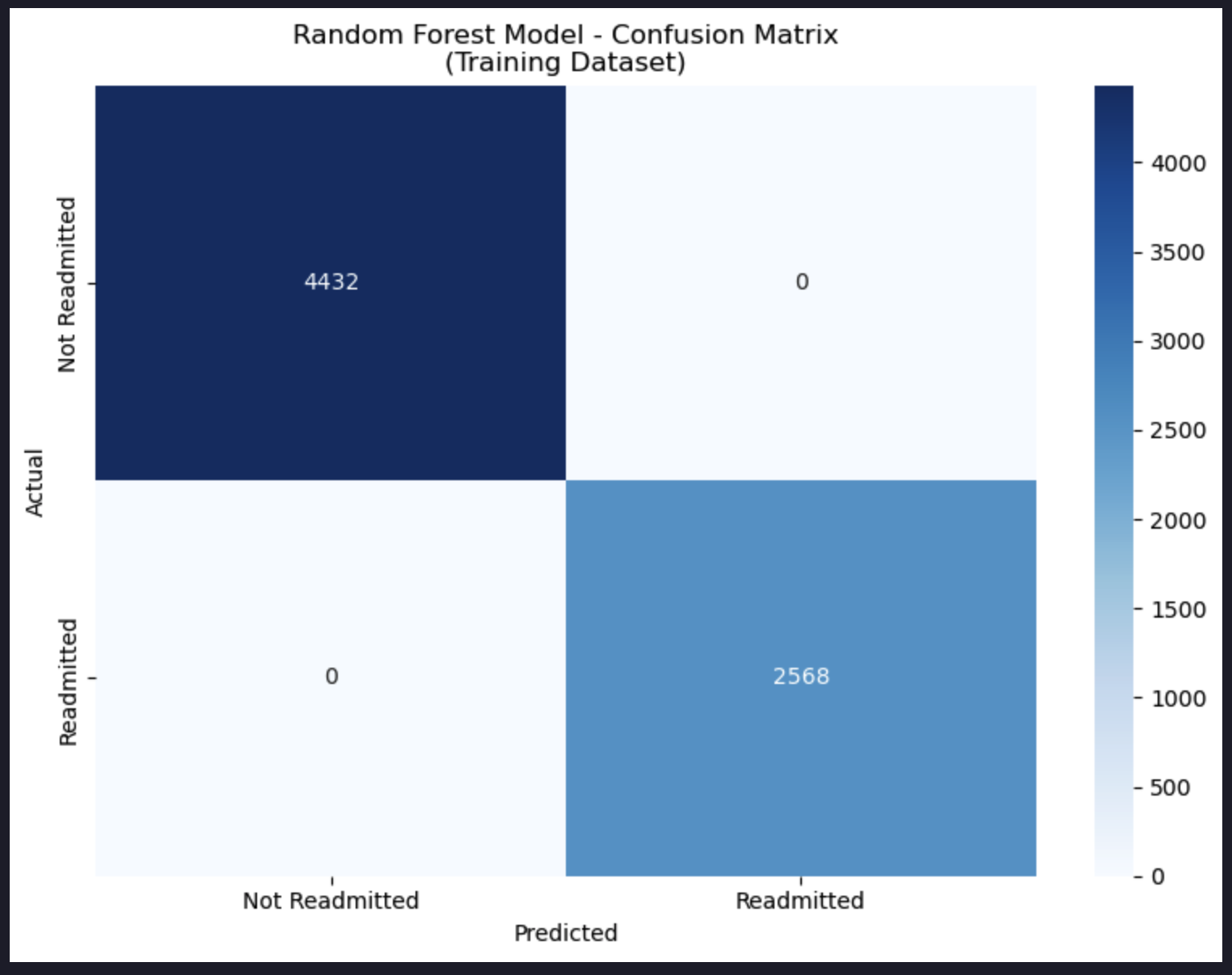




Below is a screenshot of the initial metrics as scored on the training data:



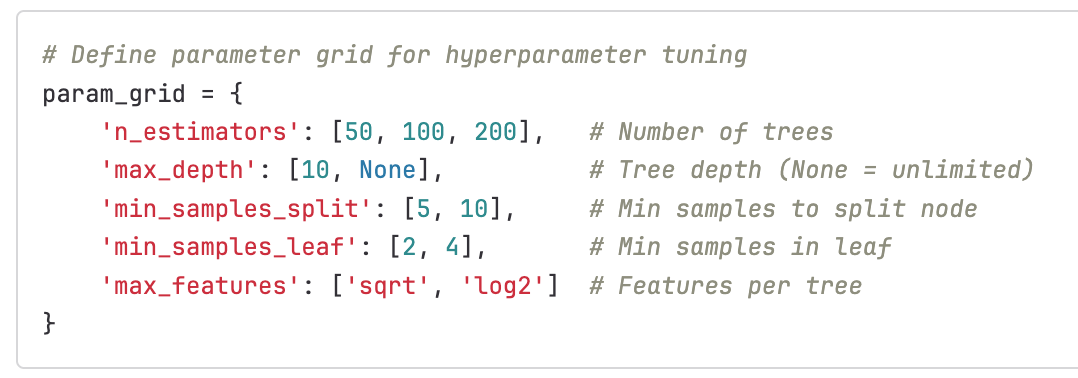




### E3. Hyperparameter Turning

**Hyperparameters Selected for Tuning:**

1. n\_estimators: Number of trees in the forest
   * More trees generally improve performance but increase computation time
   * Testing: [50, 100, 200] to balance performance vs efficiency
2. max\_depth: Maximum depth of individual trees
   * Controls overfitting - deeper trees can memorize training data
   * Testing: [10, None] to find optimal complexity
3. min\_samples\_split: Minimum samples required to split internal node
   * Higher values prevent overfitting by requiring more samples per split
   * Testing: [5, 10] from very flexible to more conservative
4. min\_samples\_leaf: Minimum samples required at leaf node
   * Controls overfitting by ensuring leaves have sufficient samples
   * Testing: [2, 4] to balance model complexity
5. max\_features: Number of features considered for best split
   * Controls randomness and prevents overfitting
   * Testing: ['sqrt', 'log2'] - standard Random Forest approaches



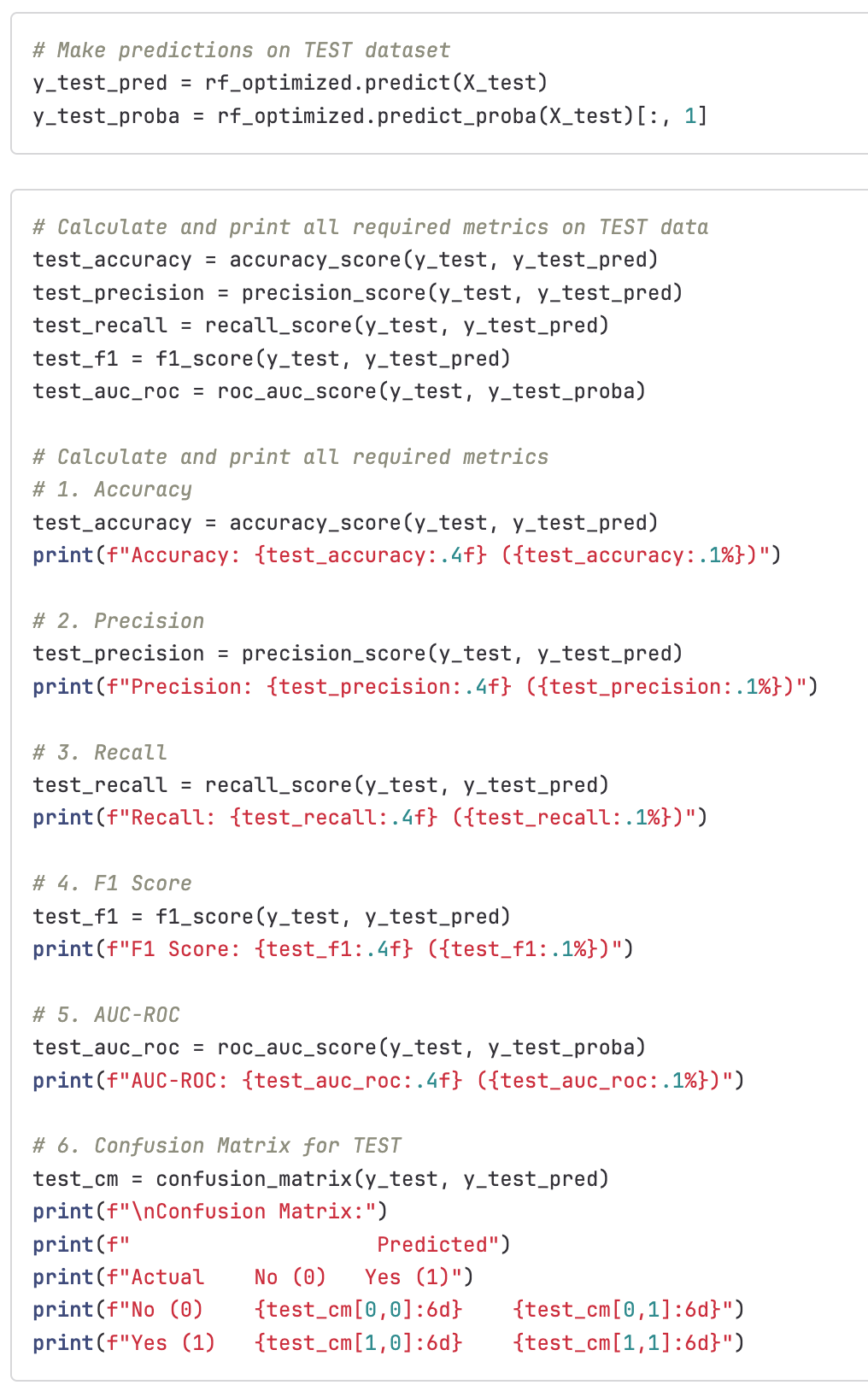
These hyperparameters were chosen because they directly control:

* Model complexity and overfitting (max\_depth, min\_samples\_split, min\_samples\_leaf)
* Ensemble diversity and performance (n\_estimators, max\_features)
* Computational efficiency vs accuracy trade-offs
* Are most impactful for Random Forest performance in medical prediction

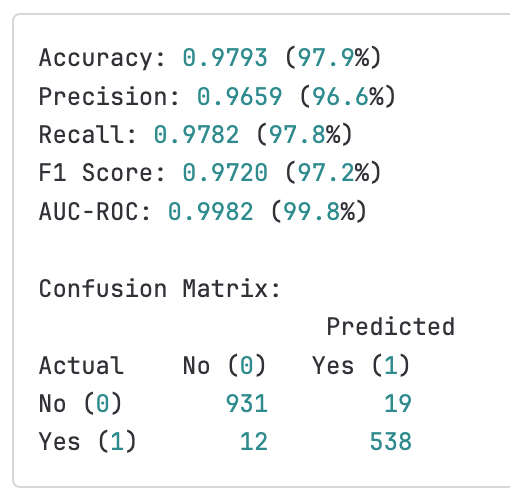


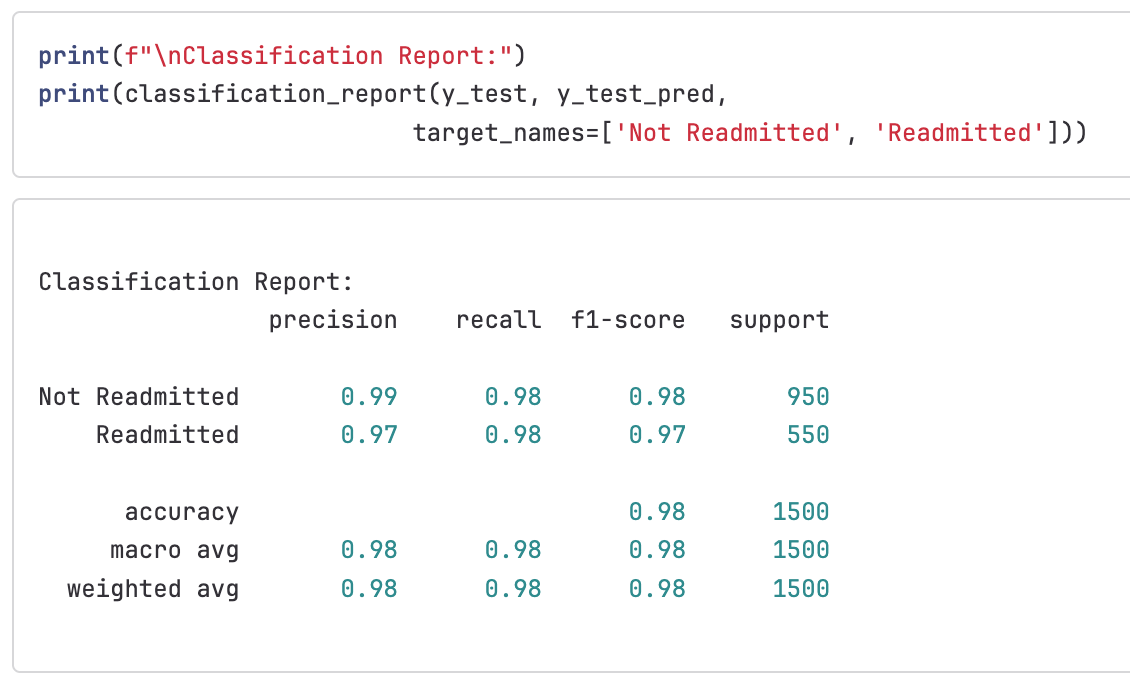


### E4. Predictions

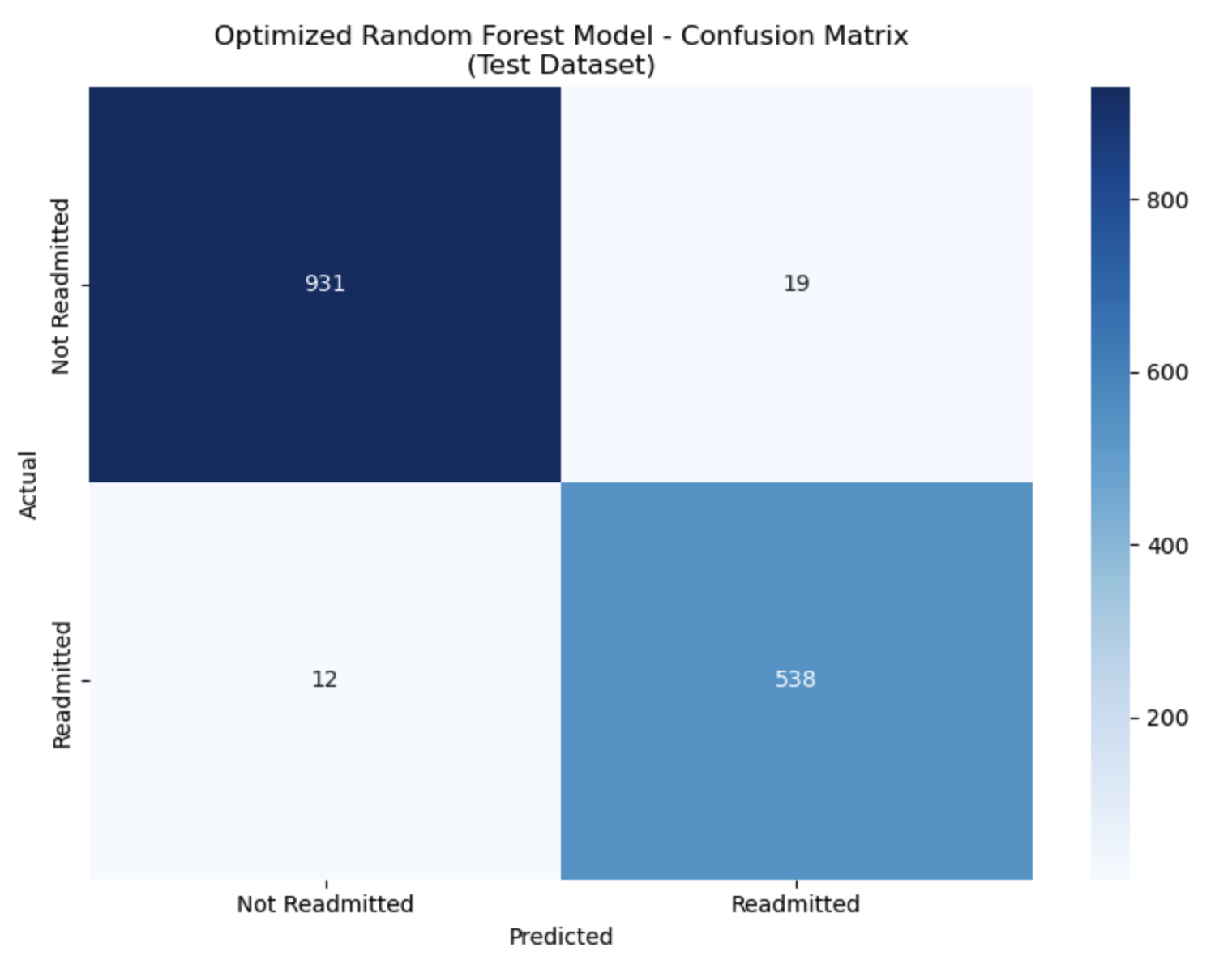


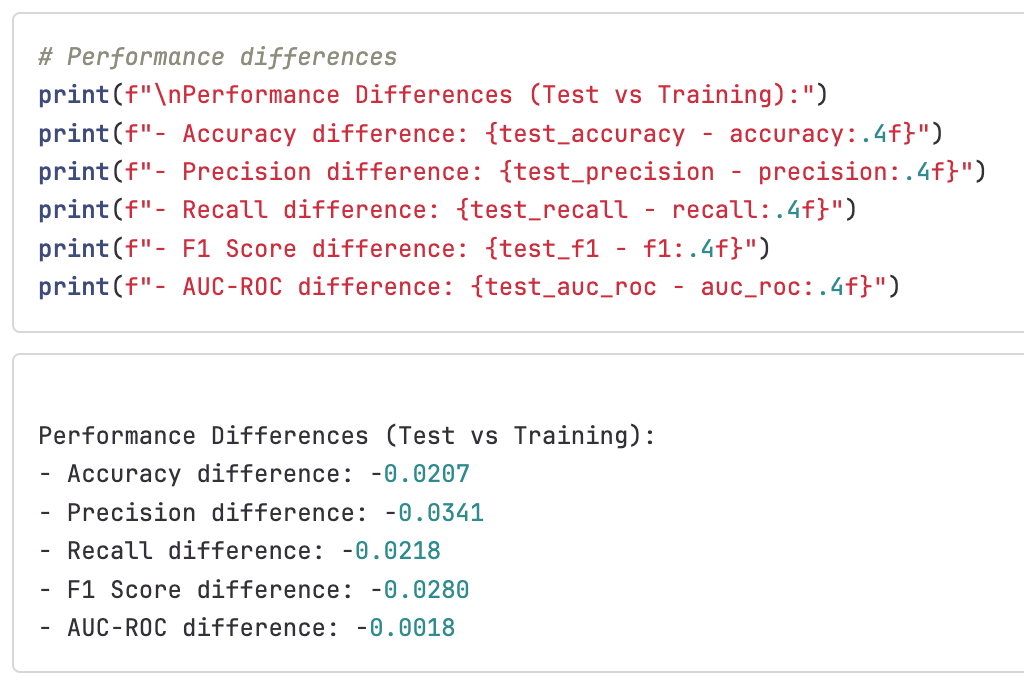
Below is a screenshot of the optimized metrics as scored on the testing data:



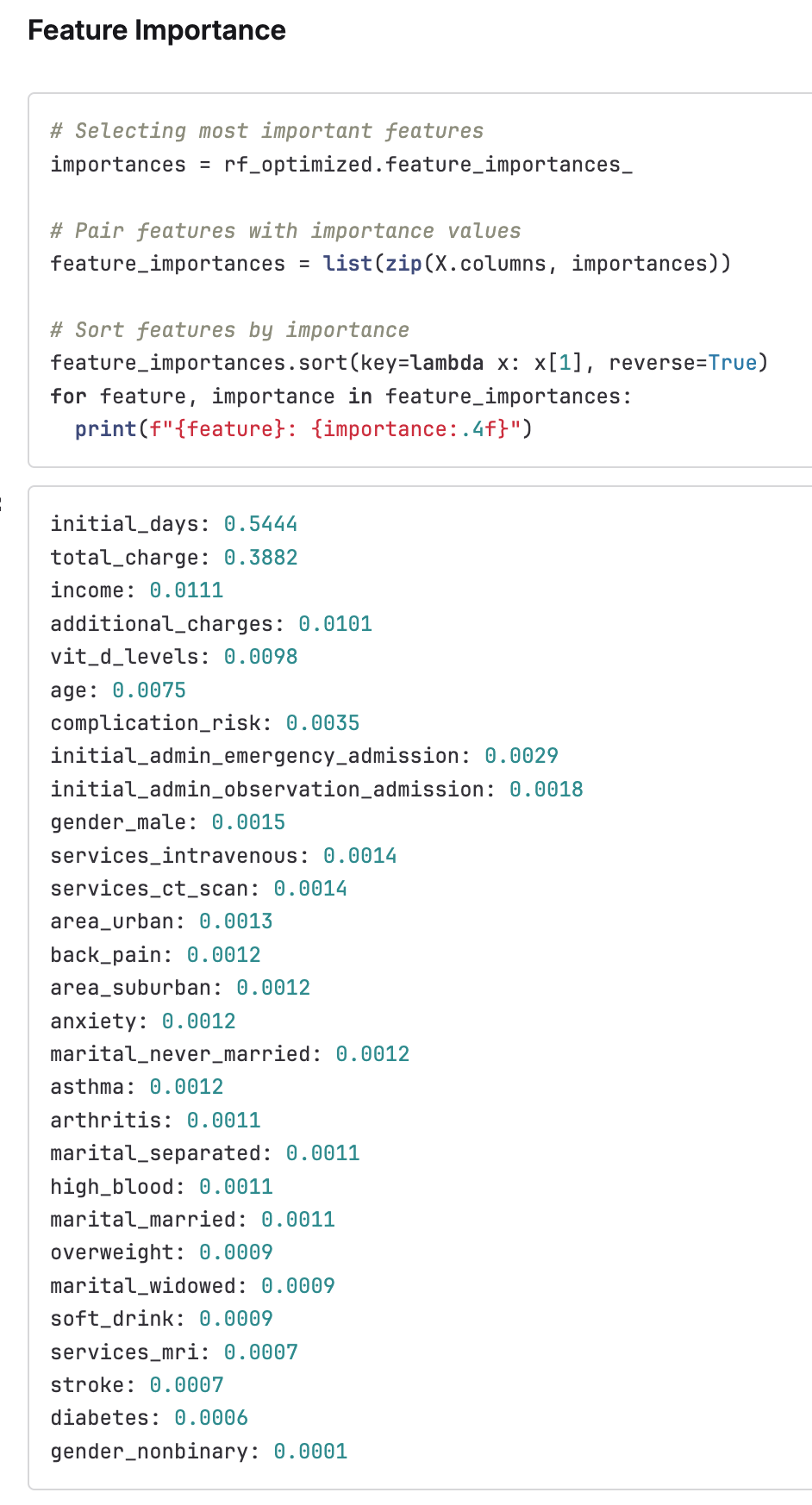








Below is a screenshot of the output of comparing feature importance:



## F. Analysis Summary

### F1. Model Evaluation

The comparison between the initial model performance on training data and the optimized performance model on test data revealed several key insights about the Random Forest Classifier.

The initial model, trained on the training dataset in Part E2, achieved perfect performance across all metrics, receiving a 100% on accuracy, precision, recall, F1 Score, and AUC-ROC. This indicates the model successfully learned patterns in the training dataset, but overfitting occurred. Perfect scores are both expected and normal for Random Forest Classification models when used on training data, again as a result of overfitting.

When splitting the data in Part E1, I took careful measure to use the stratify=y parameter when splitting the initial dataset into training, validation, and testing splits in order to keep the appropriate proportions of readmitted and non-readmitted patients for accurate testing of the model's predictive powers.

The optimized model was developed in Part E4 using the validation training dataset and then used on the testing dataset. All metrics performed at a 96% or above, showing that the hyperparameter tuning solved for overfitting while representing the model's true predictive capabilities on unseen data. The slight drop in performance again suggests that the initial model was suffering from overfitting.

In the hyperparameter tuning state (Part E3) used 5-fold cross-validation on the validation dataset to find optimal parameters, balancing model complexity and generalization of its predictive capabilities. Optimization of the F1 Score ensured balanced precision and recall for medical applications.

The small performance drop between the initial model and the optimized model is normal and expected when moving from training to testing datasets. The hyperparameter tuning using k-fold cross-validation successfully optimized the model for true predictive performance rather than the over-estimations that the overfitting was causing. The AUC-ROC scores indicate good discrimination ability between readmitted and non-readmitted patients and the optimized model generalizes well to unseen data despite the small drop in metric scores. This results in a better identification of true positives and fewer false positives.

### F2. Results and Implications

The optimized Random Forest Classification model successfully predicts patient readmission risk with approximately 96.9% accuracy, providing actionable insights for reducing patient readmission and building or improving upon hospital intervention programs.

**Clinical Implications:**

1. Early risk identification: The model enables early identification of high-risk patients before discharge, allowing medical staff to implement targeted interventions.
2. Resource allocation: Hospitals can prioritize limited resources toward patients most likely to be readmitted.
3. Personalized care plans: Feature importance rankings reveal which patient characteristics most strongly predict readmission, enabling personalized care strategies.

**Feature Importance Insights:**

The most significant features demonstrate that length of initial hospital stay (initial\_days) and total charges (total\_charges) are key predictors, likely due to longer stays and higher charges indicating more severe and/or more involved treatment needs.

**Business Impact:**

1. Cost reduction: Preventing readmissions reduces CMS penalties and associated costs.
2. Improved patient outcomes: Targeted interventions improve patient health and satisfaction.
3. Operational efficiency: Predictive insights enable better discharge planning and resource management.

**Healthcare Quality:**

The model supports evidence-based medicine by providing data-driven insights into readmission risk factors, ultimately improving the quality of patient care.

### F3. Limitation

Perhaps the biggest limitation of this particular analysis is that the model was trained and tested on data from a single hospital chain, limiting its generalizability to other healthcare systems with different patient populations, treatment approaches, or geographic locations.

Factors such as regional health disparities, varying socioeconomic conditions, patient access to hospitals and healthcare centers, and different hospital protocols could significantly impact the model's performance when applied to other healthcare organizations or hospital chains.

The model would need to be validated before deployment using different healthcare settings to ensure continued accuracy and relevance.

### F4. Course of Action

The original question asked is: Can we predict which patients are most likely to be readmitted within 30 days of discharge so that our hospital chain can implement targeted intervention programs to improve patient outcomes and reduce readmission penalties?

My initial recommended course of action is to focus on high-risk predictor variables like length of initial stay (measured in days) and total charge.

I would recommend deploying the model's use in 2-3 test hospitals and determining if the predictions can be generalized beyond the initial dataset provided.

I would also recommend training discharge planning teams and patient case managers on interpreting the model predictions and feature importance scores; this can provide flags on which patients are most at risk of readmission.

Finally, after further testing, I would recommend integrating the model with the electronic health records (EHR) system to automatically generate risk scores during patient stays.

With these things in mind, further courses of action that can be rolled out over time include developing standardized protocols for high-risk patients, establishing dedicated care coordination teams, hospital chain-wide rollout, performance monitoring to check for intervention effectiveness, and financial impact tracking.

Over the long term, model maintenance will need to be addressed to account for performance degradation and retraining, a consideration for broader model expansion to predict other patient outcomes, and improved financial impact tracking to determine return on investment through reduced penalties and improved patient satisfaction.

## G. Panopto Recording

​​<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=718e590c-a72a-4279-84d8-b33200208e80>

## H. Sources for Third-Party Code

Codecademy (n.d.) *Encoding Categorical Variables*. <https://www.codecademy.com/journeys/data-scientist-ml/paths/dsmlcj-22-machine-learning-foundations/tracks/dsmlcj-22-machine-learning-fun/modules/dsml-data-transformations-for-feature-analysis-2f0d46b3-b281-4d68-a8ac-30a198aae004/articles/fe-encoding-categorical-variables>

Codecademy (n.d.). *Training, Validation and Test Dataset*.<https://www.codecademy.com/journeys/data-scientist-ml/paths/dsmlcj-22-machine-learning-i/tracks/dsmlcj-22-supervised-learning-i/modules/mle-evaluation-metrics-98226eca-cdc3-42e7-98ee-9336cdb65a8c-a92db6b9-8ef8-4381-95d3-1e806194d196/articles/mlfun-training-test-validation>

## I. Sources

WGU Course Materials