Predictive Analytics for Patient Discharge Using Electronic Health Records

MSc Research Project
MSCDAD_JAN24A_O

Shrey Sanjay Kekade
Student ID: 23194316

School of Computing
National College of Ireland

Supervisor: Dr David Hamill

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Shrey Sanjay Kekade		
Student ID:	23194316		
Programme:	MSc Data Analytics	**Year:**	2024
Module:	MSc Research Project		
Lecturer:	Dr David Hamil		
Submission Due Date:	12/12/2024		
Project Title:	Predictive Analytics for Patient Discharge Using Electronic Health Records		
Word Count:	893	**Page Count:**	7

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Shrey Sanjay Kekade

Date: 10/12/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Shrey Sanjay Kekade
Student ID: 23194316

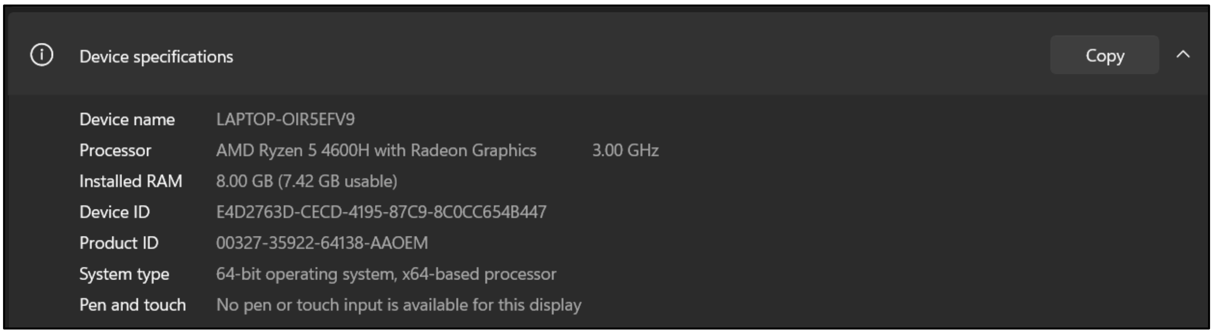
1. Introduction

This manual provides detailed instructions to replicate the analysis conducted in this research project. It includes the hardware and software requirements, data and directory setup, and guidance on running the provided notebooks. This ensures the reproducibility of results and supports further exploration of the project's methodology.

2. Hardware Requirements

The project was executed on the following system specifications:

The system used for these experiments is outlined in Figure 1, which provides a detailed summary of the hardware specifications. The operating system used is *Windows 11 Home Single Language (Version 23H2)*. Ensure the system has similar or better specifications to handle the data processing and modeling efficiently.



Device name	LAPTOP-OIR5EFV9
Processor	AMD Ryzen 5 4600H with Radeon Graphics 3.00 GHz
Installed RAM	8.00 GB (7.42 GB usable)
Device ID	E4D2763D-CECD-4195-87C9-8C0CC654B447
Product ID	00327-35922-64138-AAOEM
System type	64-bit operating system, x64-based processor
Pen and touch	No pen or touch input is available for this display

Figure 1: Device Specifications

For optimal performance, it is recommended to use a system with similar or higher specifications to handle data processing and model training efficiently.

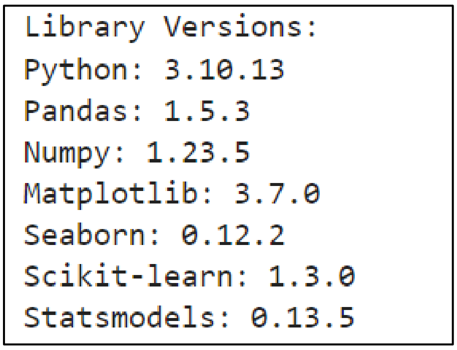
3. Software Requirements

3.1 IDE and Environment

- Environment Management: *Anaconda Navigator* (Conda Version 23.3.1).
- Integrated Development Environment (IDE): *JupyterLab* (Version 4.1.6) and *Jupyter Notebook* (Version 7.0.6), both included with Anaconda.

3.2 Python and Libraries

The analysis was conducted using *Python 3.10.13* within the Jupyter Notebook environment. Install the following libraries to ensure compatibility:

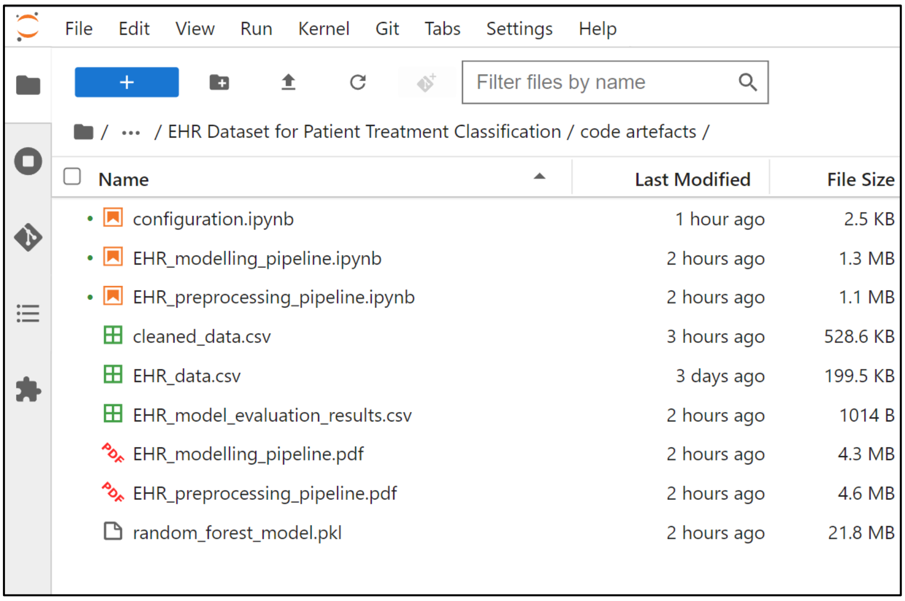


```
Library Versions:
Python: 3.10.13
Pandas: 1.5.3
Numpy: 1.23.5
Matplotlib: 3.7.0
Seaborn: 0.12.2
Scikit-learn: 1.3.0
Statsmodels: 0.13.5
```

Figure 2: Library Versions

3.3. Directory Setup

Create the following directory structure to organize the files:



/ ··· / EHR Dataset for Patient Treatment Classification / code artefacts /

Name	Last Modified	File Size
configuration.ipynb	1 hour ago	2.5 KB
EHR_modelling_pipeline.ipynb	2 hours ago	1.3 MB
EHR_preprocessing_pipeline.ipynb	2 hours ago	1.1 MB
cleaned_data.csv	3 hours ago	528.6 KB
EHR_data.csv	3 days ago	199.5 KB
EHR_model_evaluation_results.csv	2 hours ago	1014 B
EHR_modelling_pipeline.pdf	2 hours ago	4.3 MB
EHR_preprocessing_pipeline.pdf	2 hours ago	4.6 MB
random_forest_model.pkl	2 hours ago	21.8 MB

Figure 3: Directory Setup

Jupyter Notebooks:

- *configuration.ipynb*: Contains project configuration instructions.
- *EHR_preprocessing_pipeline.ipynb*: Handles data preprocessing, feature engineering, and preparation. Creates the *cleaned_data.csv* file.
- *EHR_modelling_pipeline.ipynb*: Manages model training, tuning, and evaluation. Creates the EHR_model_evaluation_results.csv and random_forest_model.pkl files.

Data Files:

- *EHR_data.csv*: Raw dataset.
- *cleaned_data.csv*: Processed dataset ready for modeling.

Results and Artifacts:

- *EHR_model_evaluation_results.csv*: Contains model evaluation metrics.
- *random_forest_model.pkl*: Saved Random Forest model for reuse.

Documentation:

- *EHR_modelling_pipeline.pdf*: PDF version of the modeling pipeline notebook.
- *EHR_preprocessing_pipeline.pdf*: PDF version of the preprocessing pipeline notebook.

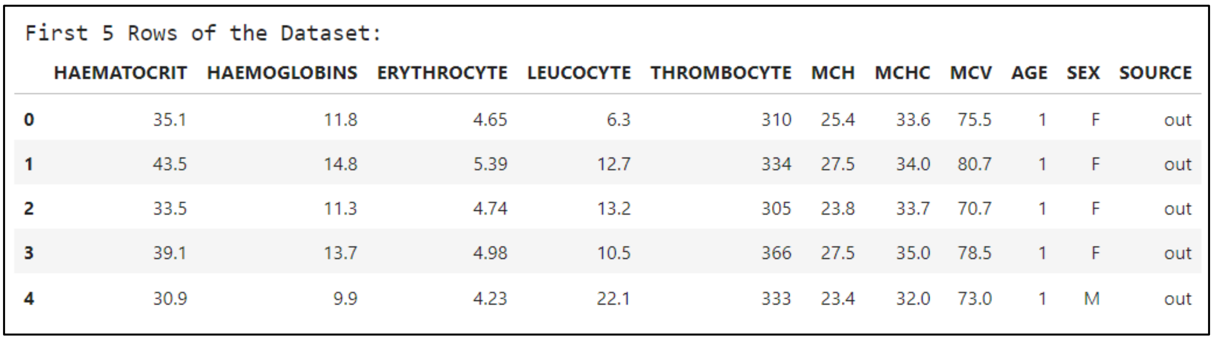
Ensure all files are placed in the appropriate directories, as shown in the figure, to facilitate reproducibility.

4. Dataset

The dataset used in this project is the Electronic Health Record Predicting dataset, which was collected from a private hospital in Indonesia. It includes laboratory test results from patients and is used to predict the next treatment classification: whether a patient requires in-care treatment or can be discharged for out-care.

Dataset Details:

- Source: Mendeley data https://doi.org/10.17632/7kv3rctx7m.1 .
- Type: Tabular data
- Task: Classification prediction to determine patient treatment outcomes.
- Structure:
  - Laboratory test results as features.
  - Target variable ‘*SOURCE’* indicating in-care or out-care classification.



First 5 Rows of the Dataset:

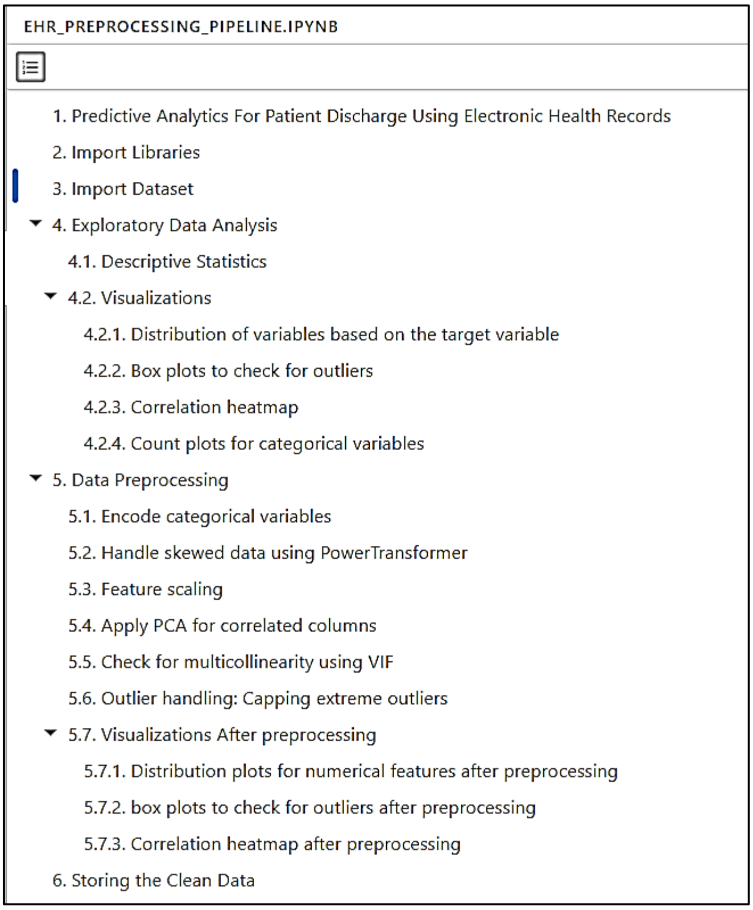
	HAEMATOCRIT	HAEMOGLOBINS	ERYTHROCYTE	LEUCOCYTE	THROMBOCYTE	MCH	MCHC	MCV	AGE	SEX	SOURCE
0	35.1	11.8	4.65	6.3	310	25.4	33.6	75.5	1	F	out
1	43.5	14.8	5.39	12.7	334	27.5	34.0	80.7	1	F	out
2	33.5	11.3	4.74	13.2	305	23.8	33.7	70.7	1	F	out
3	39.1	13.7	4.98	10.5	366	27.5	35.0	78.5	1	F	out
4	30.9	9.9	4.23	22.1	333	23.4	32.0	73.0	1	M	out

Figure 4: Dataset Overview

5. Workflow Instructions

5.1 Preprocessing Pipeline

The *EHR_preprocessing_pipeline.ipynb* notebook performs data cleaning, feature engineering, exploratory data analysis, and prepares the dataset for modeling. The detailed workflow structure is illustrated in Figure 5. Key steps include handling missing values, transforming skewed data, performing dimensionality reduction, and addressing multicollinearity.



EHR_PREPROCESSING_PIPELINE.IPYNB

1. Predictive Analytics For Patient Discharge Using Electronic Health Records
2. Import Libraries
3. Import Dataset
4. Exploratory Data Analysis
    4.1. Descriptive Statistics
    4.2. Visualizations
        4.2.1. Distribution of variables based on the target variable
        4.2.2. Box plots to check for outliers
        4.2.3. Correlation heatmap
        4.2.4. Count plots for categorical variables
5. Data Preprocessing
    5.1. Encode categorical variables
    5.2. Handle skewed data using PowerTransformer
    5.3. Feature scaling
    5.4. Apply PCA for correlated columns
    5.5. Check for multicollinearity using VIF
    5.6. Outlier handling: Capping extreme outliers
    5.7. Visualizations After preprocessing
        5.7.1. Distribution plots for numerical features after preprocessing
        5.7.2. box plots to check for outliers after preprocessing
        5.7.3. Correlation heatmap after preprocessing
6. Storing the Clean Data

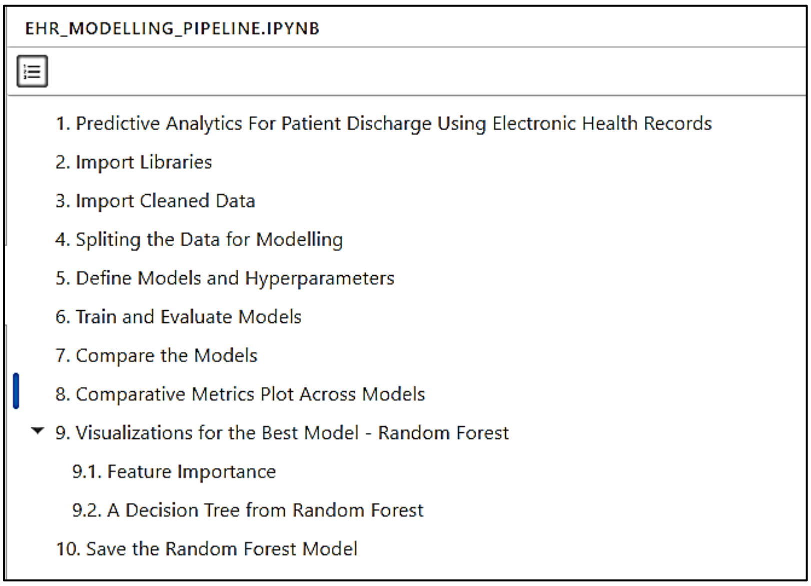
Figure 5: EHR_preprocessing_pipeline.ipynb Outline

Steps:

- Open the notebook in *JupyterLab*
- Ensure *EHR_data.csv* is present in the directory.
- Execute all cells sequentially.
- The processed dataset will be saved as *cleaned_data.csv* in the directory.

5.2 Modeling Pipeline

The *EHR_modelling_pipeline.ipynb* notebook trains and evaluates machine learning models for patient discharge prediction. The workflow, as outlined in Figure 6, includes data splitting, hyperparameter tuning, model evaluation, and comparison. Additional visualizations are provided for the Random Forest model, including feature importance and decision tree plots.



EHR_MODELLING_PIPELINE.IPYNB

1. Predictive Analytics For Patient Discharge Using Electronic Health Records
2. Import Libraries
3. Import Cleaned Data
4. Spliting the Data for Modelling
5. Define Models and Hyperparameters
6. Train and Evaluate Models
7. Compare the Models
8. Comparative Metrics Plot Across Models
9. Visualizations for the Best Model - Random Forest
   - 9.1. Feature Importance
   - 9.2. A Decision Tree from Random Forest
10. Save the Random Forest Model

Figure 6: EHR_modelling_pipeline.ipynb Outline

Steps:

- Ensure *cleaned_data.csv* is present in the directory.
- Open the notebook in *JupyterLab*.
- Run all cells sequentially to train and evaluate models.
- Outputs, including *model_evaluation_results.csv* and *random_forest_model.pkl*, will be saved in the directory.

6. Conclusion

This configuration manual outlines the hardware, software, and step-by-step instructions needed to replicate the project's workflows. By providing clear guidance for preprocessing and modeling pipelines, it ensures reproducibility and supports further exploration of predictive analytics in healthcare.