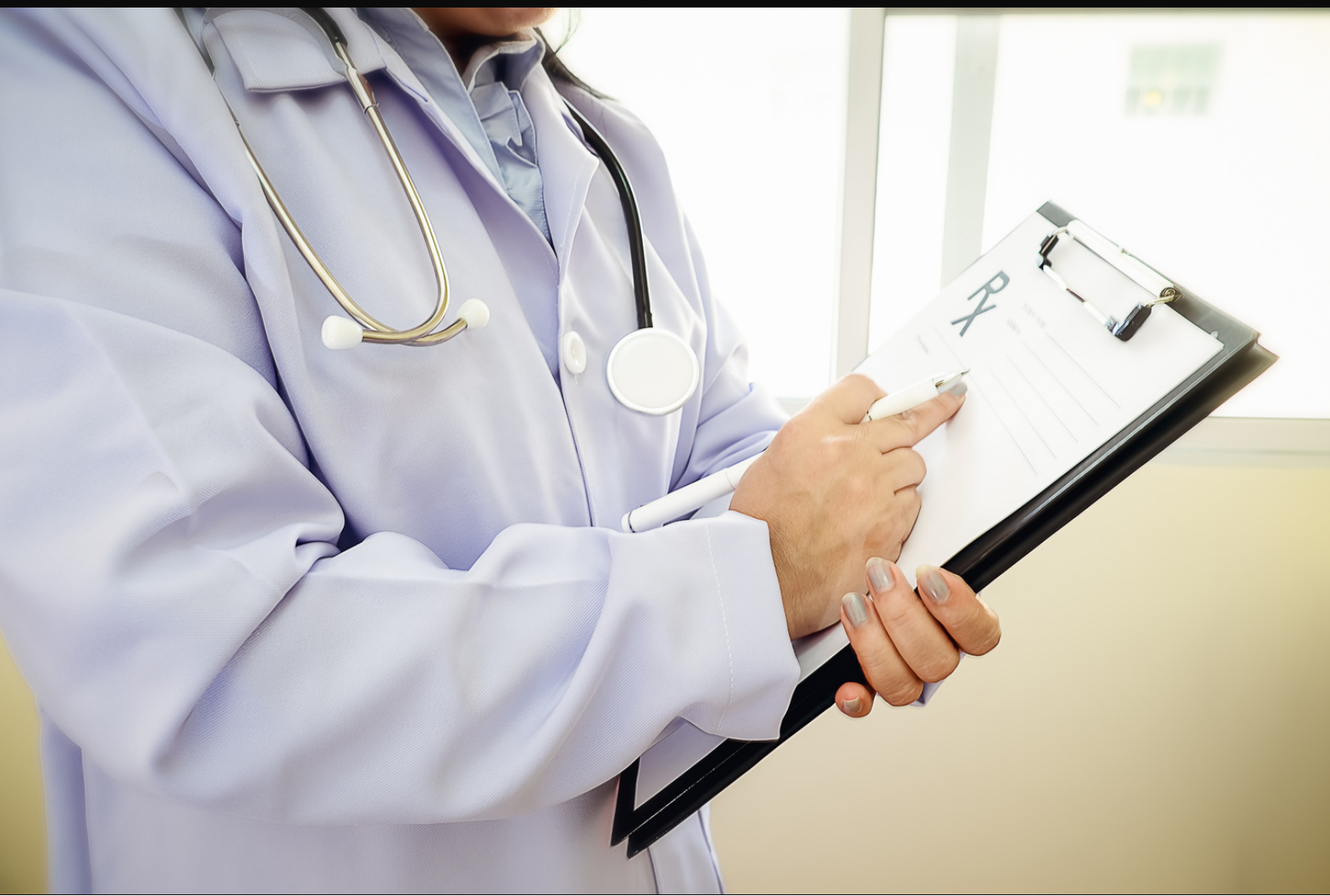
OCTOBER, 2023

READY TO BE DISCHARGED: EXAMINING HOSPITAL READMISSIONS

GROUP PROJECT
MACHINE LEARNING 2023/2024



01

I. INTRODUCTION

Hospital readmissions represent a significant challenge in the healthcare sector, both as an indicator of care quality and a driver of escalating costs. When a patient is re-admitted to the hospital within a short period after discharge, it not only indicates potential gaps in care but also adds to the financial burden on the healthcare system. In particular, readmissions of diabetic patients have been noted to contribute significantly to these costs. Therefore, being able to predict such readmissions can lead to improved patient care and substantial cost savings.

II. PROJECT GOALS

The goal of your project is two-fold:

1. **Binary Classification:** Create a classification model that can accurately predict if a patient will be readmitted to the hospital within 30 days of being discharged. A robust prediction can enable healthcare providers to implement preventive measures and provide timely intervention, potentially saving millions of dollars in healthcare costs.
2. **Multiclass Classification:** The second objective is to develop a multiclass classifier that predicts the timeframe of a patient's readmission, with the classes being "*No*", "*<30 days*", "*>30 days*". This model can provide more nuanced insights into patient risk levels and help hospitals tailor their post-discharge care and follow-up procedures accordingly.

Through these predictive models, we aim to equip healthcare providers with valuable tools that can enhance patient care quality, reduce readmission rates, and contribute to more sustainable healthcare spending.

III. DATASET

You have access to two different datasets:

In the **training set**, you will find the features and two specific ground truths associated with each encounter. Use the training data and its features to build and validate your machine-learning models. The goal will be to use the models you created and make predictions on unseen data (i.e. your test set). **Important note: You should not consider any of the target variables as features for any of the predictive models.**
In the test set, you will still have access to the same descriptive attributes associated with each encounter. However, you will not have access to the target variables for either the binary nor the multiclass problems.

The available data contains the following attributes:

ATTRIBUTE	DESCRIPTION
encounter_id	Unique identifier of the encounter
country	country
patient_id	Identifier of the patient
race	Patient's race
gender	Patient's gender
age	Patient's age bracket
weight	Patient's weight
payer_code	Code of the health insurance provider (if there is one)
outpatient_visits_in_previous_year	Number of outpatient visits (visits made with the intention of leaving on the same day) the patient made to the hospital in the year preceding the encounter
emergency_visits_in_previous_year	Number of emergency visits the patient made to the hospital in the year preceding the encounter
inpatient_visits_in_previous_year	Number of inpatient visits (visits with the intention to stay overnight) the patient made to the hospital in the year preceding the encounter
admission_type	Type of admission of the patient (e.g. Emergency, Urgent, etc...)

03

ATTRIBUTE	DESCRIPTION
medical_specialty	Medical specialty on which the patient was admitted
average_pulse_bpm	Average pulse of the patient during their stay in the hospital in beats per minute
discharge_disposition	Destination given to the patient after being discharged
admission_source	Source of the patient before being admitted in the current encounter
length_of_stay_in_hospital	Number of days between admission and discharge
number_lab_tests	Number of lab tests performed during the encounter
non_lab_procedures	Number of non-lab procedures performed during the encounter
number_of_medications	Number of distinct types of medication administered during the encounter
primary_diagnosis	Primary diagnosis (coded as first three digits of ICD9)
secondary_diagnosis	Secondary diagnosis (first three digits of ICD9)
additional_diagnosis	Additional secondary diagnosis (first three digits of ICD9)
number_diagnoses	Number of diagnoses entered to the system
glucose_test_result	Range of the glucose test results or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured
a1c_test_result	Range of the A1C test results or if the test was not taken. Values: “>8” if greater than 8%, “>7” if greater than 7% but less than 8%, “normal” if less than 7%, and “none” if not measured.
change_in_meds_during_hospitalization	Indicates if there was a change in diabetic medications (dosage or generic name). Values: “change” and “no change”
prescribed_diabetes_meds	Yes if patient has diabetes medication perscribed. No otherwise.
medication	List containing all generic names for the medications perscribed to the patient during the encounter. Empty list if no medication was perscribed.
readmitted_binary	**Binary target:** Yes if patient was readmitted in less than 30 days, No otherwise.
readmitted_multiclass	**Multiclass target:** “<30 days” if patient was readmitted in less than 30 days after being discharged. “>30 days if patient was readmitted to the hospital but only after more than 30 days after the current discharge. No otherwise.

indicador de controlo de diabetes

Targets

IV. OUTLINE

Your project deliverables (especially the report) should respect the following outline:

Abstract
A small summary of your work (200 to 300 words). The abstract should give an overview of your work: What is the context? What is your main hypothesis? What did you do? What were your main results, and what conclusions did you draw from them?

I. Introduction
- Overview of the project
- Main goals of the project (being a requirement for the course does not count)
- Did you find any research with similar objectives? What has been done? What did other researchers find? What would you expect your results to be based on their previous findings?

II. Data Exploration and Preprocessing
- Description of data received
- Steps taken to clean and prepare the data

III. Binary Classification
- Additional preprocessing steps adopted
- Description of the actions taken
- Results and discussion of main findings

IV. Multiclass Classification
- Additional preprocessing steps adopted
- Description of the actions taken
- Results and discussion of main findings

V. Conclusion
- Summary of the findings
- Do the findings match what you initially expected? How?
- Discussion of limitations of your work (e.g. what could you have done differently)
- Suggestions for possible work to follow on your work.

V. DELIVERABLES

Upon the project's deadline, you will be required to submit:

- A Jupyter notebook (or a zip of multiple notebooks) featuring all the code you used throughout the project to:
  a. Decide on your final solution
  b. Obtain your final results (code that helped you make decisions but does directly contribute to reach should be included, but commented).
- A report that describes the analytical processes and the conclusions obtained with, at most, 10 pages (excluding cover, abstract and annexes).
  - The file naming format should follow ***Machine_Learning_GroupXX_Report.pdf***, where ***GroupXX*** should be your group number. The report should follow these settings**:**
    - **Heading 1: Calibri, Size 14 pt, in bold**
    - **Heading 2 (if needed): Calibri, Size 13 pt, in bold**
    - **Text: Calibri, Size 11 pt, line spacing of 1.15 pt and paragraph spacing of 6 pt**
  - The body of text should only include Figures and Tables that are essential to understand your work. Supporting figures and Tables can be added to Annexes.
  - Please make sure all figures and Tables (including the ones in annexes) are identified and referenced in the text. Any figure or table should have an explicit purpose to be included.

VI. EVALUATION

Your work will be evaluated according to the following criteria:

CRITERIA	PERCENTAGE (%)	MAXIMUM GRADE (OUT OF 20)
Report Quality and Storytelling	15	3
Data Exploration	10	2
Initial data preprocessing	15	3
Binary Classification	20	5
Multiclass Classification	20	4
Conclusion	10	2
Creativity and Other Self-studies	5	1

06

Your grade will reflect our assessment of the quality of your work in terms of quality of writing, clarity, conciseness, correctness and efficiency. Please find below more details about what is taken into account for each topic:

- **Report Quality and Storytelling (3v):** A good report should, by itself, give the reader a clear picture of the problem you are tasked with, the steps you took, the rationale behind those steps, your main results and your insights. When referencing a figure, ensure you direct the reader's attention to the point you want to convey. This section also encompasses the overall quality of your introduction and conclusions.
- **Data Exploration (2v):** Describe the data and extract meaningful insights that you consider helpful. Avoid adding visualizations and elements that add nothing to address the problem at hand.
- **Initial data preprocessing (3v):** This section covers the initial preprocessing of your data. In essence, it should unambiguously explain the steps and rationale behind your steps in transforming the data into data usable by your predictive models.
- **Binary Classification (5v):** Describe your strategy for the text classification objective. This section is separated into different components:
  - Kaggle Performance: 1v
  - Additional Preprocessing (includes feature selection): 1.5v
  - Modelling approach (model assessment (holdout, cross-validation, etc...), algorithms used): 1v
  - Performance assessment (choice of metrics and interpretation of results): 1.5v
- **Multiclass Classification (4v):** Describe your strategy for the multiclass classification objective. This section is separated into different components:
  - Additional Preprocessing (includes feature selection): 1v
  - Modelling approach (model assessment (holdout, cross-validation, etc...), algorithms used): 1v
  - Performance assessment (choice of metrics and interpretation of results): 2v
- **Creativity and other self-studies (1v):** This topic includes applying different techniques and aspects of creativity, such as choice of visualizations, approach or techniques used. If techniques not given during practical classes are used, you should provide a theoretical explanation for them in the annexes.

07

VII. PARTING NOTES

1. **For modelling purposes, any algorithm implementation outside the vanilla scikit-learn is explicitly off-limits. Moreover, using Lazy Predict or similar AutoML packages is also not allowed.**
2. The report will be the primary method of evaluating your work. When preparing it, remember that a reader should be able to understand your work without needing to check your notebook. We won't be able to consider any steps or results not mentioned in your report.
3. Please don't provide long theoretical explanations of topics covered in class in your report.
4. Everything featured in your report must have a clear purpose. Avoid including irrelevant/unimportant/redundant information, as the space is limited, and you will need it.
5. Trustworthiness of the information you provide is key. You should look to source information you provide from peer-reviewed journals (thus, avoid citing Medium, TowardsDataScience and similar sources).
6. Before submitting, run your notebook from the start one last time (if you used a GridSearch, you can comment this cell, but you should run the final model with the GS parameters in a different cell).
7. All the unneeded code you used to obtain your final solution **should be part of your submitted notebook, but it should be commented.**
8. We will run your Jupyter Notebooks if we have any doubts. So, please make sure we can run the notebook from start to finish in one go. Notebooks that do not fulfil this condition will be penalized.
9. The report and code will pass through a process of plagiarism and AI generation checking.
10. You must submit to the Kaggle competition to get points for that component.
11. When determining the grade for your work, there will be a comparative component between your work and the works presented by your peers.

Friendly Reminders:

1. Attendance at the defense is mandatory for approval in the project. The defense has a group component and an individual component.
2. As questions are individualized, every group member should be able to understand what was done at every step of the way.
3. **If something is good enough to be mentioned in the report, it is also good enough to know. DO NOT include techniques/algorithms/steps you cannot explain in your report: we may (and probably will) ask about them in the defense.**
4. Finished is better than perfect.