Group Project

The Smith Parasite

MACHINE LEARNING 2022/2023

1 Introduction

A new disease has recently been discovered by Dr. Smith, in England. You have been brought in to investigate. The disease has already affected more than 5000 people, with no apparent connection between them.

The most common symptoms include fever and tiredness, but some infected people are asymptomatic. Regardless, this virus is being associated with post-disease conditions such as loss of speech, confusion, chest pain and shortness of breath.

The conditions of the transmission of the disease are still unknown and there are no certainties of what leads a patient to suffer or not from it. Nonetheless, some groups of people seem more prone to be infected by the parasite than others.

2 Objective of the project

In this challenge, your goal is to build a predictive model that answers the question, "Who are the people more likely to suffer from the Smith Parasite?". With that goal, you can access a small quantity of sociodemographic, health, and behavioral information obtained from the patients.

As data scientists, your team is asked to analyze and transform the data available as needed and apply different models to answer the defined question in a more accurate way. Can you build a model that can predict if a patient will suffer, or not, from the Smith Disease?

The score of your predictions is the percentage of instances you correctly predict, using the f1 score.

3 Datasets

You have access to two different datasets:

1. The training set should be used to build your machine learning models. In this set, you also have the ground truth associated to each patient, i.e., if the patient has the disease (Disease = 1) or not (Disease = 0). Is composed by:
   train_demo.csv - the training set for demographic data and the target
   train_health.csv - the training set for health related data
   train_habits.csv - the training set for habits related data

2. The test set should be used to see how well your model performs on unseen data. In this set you don't have access to the ground truth, and the goal of your team is to predict that value (0 or 1) by using the model you created using the training set. Is composed by:
   test_demo.csv - the test set for demographic data
   test_health.csv - the test set for health related data
   test_habits.csv - the test set for habits related data

The available data contains the following attributes:

Sociodemographic Data

Attribute	Description
PatientID	The unique identifier of the patient
Birth_Year	Patient Year of Birth
Name	Name of the patient
Region	Patient Living Region
Education	Answer to the question: What is the highest grade or year of school you have?
Disease	The dependent variable. If the patient has the disease (Disease = 1) or not (Disease = 0)

Health Related Data

Attribute	Description
PatientID	The unique identifier of the patient
Height	Patient's height
Weight	Patient's weight
Checkup	Answer to the question: How long has it been since you last visited a doctor for a routine Checkup? [A routine Checkup is a general physical exam, not an exam for a specific injury, illness, or condition.]
Diabetes	Answer to the question: (Ever told) you or your direct relatives have diabetes?
High$_Cholesterol$	Cholesterol value
Blood$_Pressure$	Blood Pressure in rest value
Mental Health	Answer to the question: During the past 30 days, for about how many days did poor physical or mental health keep you from doing your usual activities, such as self-care, work, or recreation?
Physical Health	Answer to the question: Thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good to the point where it was difficult to walk?

Habits related Data

Attribute	Description
PatientID	The unique identifier of the patient
Smoking_Habit	Answer to the question: Do you smoke more than 10 cigars daily?
Drinking_Habit	Answer to the question: What is your behavior concerning alcohol consumption?
Exercise	Answer to the question: Do you exercise (more than 30 minutes) 3 times per week or more?
Fruit_Habit	Answer to the question: How many portions of fruits do you consume per day?
Water_Habit	Answer to the question: How much water do you drink per day?

4 Deliverables

1. A Jupiter notebook with all the needed code implemented to obtain the results presented in the report and to obtain the performance obtained in Kaggle. The file naming format should be ”202223_Fall_AA_GroupXX_Notebook.ipynb”, where ”GroupXX” should be your group number.

2. A report that describes the analytical processes and the conclusions obtained, with at most 8 pages. The report should be organized as follows:

    - Introduction
    - Exploration
    - Preprocessing
    - Modelling

- Assessment
- Conclusion
- References
- Annexes

The report should respect the following format:

- **Heading 1:** Arial, Size 12 pt, in bold
- **Heading 2 (if needed):** Arial, Size 11 pt, in bold and italic
- **Text:** Arial, Size 10 pt, line space of 1.5 points.
- **Margins:** The default ones in word (Top, Bottom, Left and Right as 1").

All the figures and tables should be included in the Annexes (at the end of the document) and referenced in the body text, and are not included on those 8 pages mentioned previously.
The 8 pages restriction is limited to the content of the report (do not include cover, index, references or annexes)
The reports that do not follow the specified conditions will suffer penalization on the grade.

The file naming format should be "202223_Fall_AA_GroupXX_Report.pdf", where "GroupXX" should be your group number.

4.1 Notes

- We will evaluate all the topics mentioned based on the report - a well-structured and succinct report will have a big weight on the evaluation.

- The jupyter notebook will be analyzed only if some doubt arises during the report evaluation. If some steps were done in the Jupyter notebook but not described in the report, we will not evaluate those. As an example, imagine you check the outliers, and at the end of your project, you decide to keep them. In the report, you should mention how you check if you had outliers, what the steps were to remove them and why you decide to keep them at the end, among other insights that can be relevant. The jupyter notebook should be delivered with all the cells already run.

- The report and the code will pass through a process of plagiarism checking.

- The report should clearly refer (on a cover or on the first page if no cover is included) the group number, the students' names and the students' numbers.

- **For more information, please read the Kaggle competition rules carefully.**

- **The deadline for submission of all documents is set until the end of December 23, 2022. Your report (pdf format) and your Jupiter notebook (ipynb format) should be submitted in moodle by this date. Additionally, the final submission in Kaggle must be selected. One submission per group is enough.**
  For each day of delay, there will be a discount of 1 value on the final grade. The maximum possible number of days of delay is three days (with a penalization of 3 values out of 20).

5 Evaluation Criteria

The following table quantifies the major evaluation criteria.

Criteria	Percentage	Maximum Grade
		(out of 20)
Kaggle performance	20%	4
Report-quality and Story-telling	15%	3
Exploration	10%	2
Preprocessing	10%	2
Modelling	20%	4
Performance Assessment	10%	2
Other predictive models (not given during classes)	5%	1
Creativity & Other Self-Study	10%	2
TOTAL	100%	20

A project that focus only on the techniques and methodologies approached during the practical classes will have at most 17 values. The remaining 3 values are possible to achieve if contributions based on self-study and creativity are applied, and clearly explained on the report.

This bullet-list provides some details about each aspect:

- **Kaggle performace:** The performance obtained on Kaggle, on the submission selected (F1 Score).
- **Report-quality Story-telling:** Each report should describe the steps and main insights along the process. Clarity, synthesis, objectiveness, and business-contextualization are very welcome.Your decisions and steps must be reasonably justified by the previous findings (when this is possible and feasible), your hypothesis and findings must be related to the problem's business-context, etc.

- **Exploration:** Describe the studied population using statistical measures, meaningful insights and visualizations representative of the major insights.

- **Preprocessing:** Includes all the needed steps to transform the raw data into the data prepared to model. Involves all the steps for cleaning, transform and reduce the dataset. It also involves the creation of new variables (if any) from the original input features and the explanation of those. If new variables are created, those should be mentioned and described clearly on a table (to be included in the annexes).

- **Modelling:** the implementation of different predictive algorithms and the process of fine-tuning those models. The application of additional models not given during classes are optional and considered as points in ”Other predictive models”.

- **Performance Assessment:** The comparison of different models and their performance.

- **Other predictive models:** A theoretical explanation of the algorithm should be provided in the annex (not included in the 8 pages). Involves the depth and the quality of the comparative analysis provided by the different algorithms, the theoretical explanation of the algorithm itself and the justification of the chosen parameters;

- **Creativity and Other Self-Study:** If other techniques not given during practical classes are applied, a theoretical explanation of the algorithm / technique should be provided in the annex (not included in the 8 pages). This topic includes not only the application of different techniques but also aspects of creativity, such as the the quality of visualizations, plots and others.

All topics are evaluated through a comparison of the work provided by the different groups.