

**Machine Learning I**

**Report**

**Project title**

Subtitle

**GroupID**

**Date**

**João Ferreira - 20221912**

**Rodrigo Maia - 20221934**

**Miguel Mendes - 20221904**

**Suelen Faruk – 20221997**

**João…….. – 2022………**

**NOVA Information Management School**

**Instituto Superior de Estatística e Gestão de Informação**

Universidade Nova de Lisboa

Abstract

The abstract is a paragraph containing 250-350 words summarising your report and findings. As a rule of thumb, it should be arranged in five sections: subject matter and purposes of the study (Introduction), the methodology adopted (Methodology), your results (Results), the interpretation of your results (Discussion) and a conclusion of your work (Conclusions) based on your findings. Sentences which do not convey useful information should be avoided.

While the Abstract is the first visible section of your work, it will often be the last part you write. In conclusion, an abstract should be viewed as a miniversion of the report - it should provide a brief summary of each main section [1].

Keywords

Keyword 1; Keyword 2; Keyword 3; Keyword 4

*[ You must have at least one keyword. You can enter up to 6 keywords separated by a semi-colon. Capitalize the first letter of keywords. Use full phrases rather than acronyms or abbreviations (use Machine Learning rather than ML). ]*

# 

# Introduction

This report summarizes a project to develop a predictive model that, based on information about each student, determines which students should be accepted into wizarding schools, ensuring that only the applications of the best and most complete candidates are accepted.

A dataset of 713 applications was used to develop the predictive model. It was with this data, and after studying, processing and analyzing each candidate, their characteristics and qualifications, that we built several models for the intended purpose.

This project is very important, since this model will play a crucial role in predicting which students meet the necessary conditions to enter the most prestigious schools of witchcraft. For this reason, in order to ensure that only the most deserving candidates gain admission to the realm of magical education, it is essential to make as few mistakes as possible when selecting students, which will only be possible with the development of a good model.

# Background

This section should explain the theoretical background of the techniques/algorithms that have not been explored during the practical classes and applied in the project. On the other hand, the theoretical background of the methods demonstrated in the practical classes should not be exposed in the report.

## Level 2 title

Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables.



Figure 2.1 – Illustrative figure

Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables

Table 2.1 – Illustrative table

|  |  |
| --- | --- |
| **Title** | **Title** |
| Text | Number |
| Text | Number |
| Text | Number |

Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables.

### Level 3 title

Example of an unnumbered list:

* Item 1
* Item 2
* Item 3

# Methodology

In this project, we have two datasets. One, the training dataset, which includes information about students who have previously been accepted or not into the wizardry school, as well as the final outcome of their application. The other, the test dataset, has the same variables as the training one, but without the dependent variable, that we want to predict. For the next steps, only the training dataset will be used.

A screenshot of a computer

Description automatically generatedWe started the project with the feature selection phase, with the aim of improving the efficiency and performance of our model. We began by conducting an analysis of our data to understand the characteristics of our dataset. At this stage, when we realized that the school dormitory variable had more than 70% missing values (76%), we decided to remove it, according to the rule of thumb described in <https://medium.com/@danberdov/dealing-with-missing-data-8b71cd819501>.

A screenshot of a computer

Description automatically generatedThe first step in feature selection was to divide our dataset into training and validation sets, and we did this using 70% for training, 30% for validation, using a random state of 15 and stratifying the target. After this step, we began the feature selection itself, starting with the categorical variables (Program, Student Gender, School of Origin, Favourite Study Element, Study Element). We did this using the chi-square test of independence and concluded that the Favourite Study Element variable was not relevant to our predictions, so it was discarded.

Having completed the feature selection of the categorical variables, we began selecting the remaining variables (numerical and categorical that had not previously been excluded). We start by transforming all the non-numerical variables into numerical ones by converting them to dummy variables. In addition, as we had some variables with missing values, we used the KNNImputer method to insert values and thus have 100% complete observations. After applying the last two processes (dummy variables and KNN imputer) to the training and test sets, we repartitioned the resulting training dataset, using, as before, a train size of 70%, a random state of 15 and also stratifying the target.

A graph of different colored rectangular shapes

Description automatically generated with medium confidenceAt this point, using a multi-layer perceptron classifier model, we tested the best way to scale our features so that we could continue with the feature selection process. We tested the scores obtained for the validation sets using 5 different techniques: no scale at all, minmax scaler with default range [0,1], minmax scaler with range [-1,1], standard scaler and robust scaler. After analysing the results, we concluded that the best way to scale our features was minmax with range [-1,1].

A table of numbers and text

Description automatically generatedWith this information, we scaled our variables in analysis and continued the process. We tested their univariance, concluding that none was univariate, and after that we tested the correlation between all the variables (independent and target), concluding that the pairs Sorcery School - Magi Academy, Financial Background - Sorcery school and Financial Background - Magi Academy had significant correlation. In the same way we verified that the variable Male had considerable correlation with the target variable.

To finish this phase, we used 3 different methods to evaluate the importance of our features: Recursive Feature Elimination on a Logistic Regression model, Lasso Regression and the Gini and Entropy concepts in a Decision Tree Classifier. Analysing the results of RFE, we concluded that using 9 variables would give the best performance, and so the only variable excluded would be Mystic Academy. Then, and according to the next method, Lasso, all variables seemed good predictors except Student Family. Eldertree Enclave, Financial Background and Mystic Academy are doubtful variables. Lastly, according to the decision tree (with gini and entropy criterion) all variables seemed good estimators, although Students Family, Mystic Academy, Eldertree Enclave and Magi Academy are the weaker ones (and can eventually be excluded). Having all this analysis in mind, and after combining the results from the 4 methods we utilized (Spearman Correlation, RFE on Logistic Regression, Lasso Regression and Decision Tree Classifier with entropy and gini criteria), our final decision was to drop the variables Student Family and Mystic Academy, as this two were consensually classified as not very relevant for A white background with black text

Description automatically generatedour predictions.

Algorithms:

After the feature selection phase, we went on to implement various classification algorithms in order to find the most suitable model. Before we started developing our models, we checked whether the dataset was balanced or not, to understand which metrics were most important when evaluating them. It was found that it was not, with 65% of applications not accepted and 35% of applications accepted.

At this stage, the following models were worked on: Logistic Regression, KNN Classifier, Gaussian Naïve Bayes, Decision Tree, Bagging Classifier with either Decision Tree or KNN as base estimator, Random Forest Classifier, Stacking Classifier (tested with different combinations of algorithms), AdaBoost Classifier, Gradient Boost Classifier and MLP Classifier. For each algorithm mentioned above, its parameters were optimized through a grid search using some predefined hyperparameters. A scoring function was also used to help identify which combination of these hyperparameters generated the best results. It was this process that allowed us to tune each algorithm and maximize its predictive ability.

After training each model, the results were checked for each one using the classification report and the confusion matrix. The classification report allowed us to evaluate some metrics such as precision, recall and F1 score and thus understand the performance of the model under evaluation. The confusion matrix, in turn, showed us the count of true positives, true negatives, false positives and false negatives, contributing to a better interpretation of the results obtained.

This repetitive process allowed us to compare the strengths and weaknesses of each algorithm and thus ensure that we chose the model which, given the proposed objective, would have both strong performance and considerable interpretability.

Finally, and although the results obtained by each algorithm will be explained later in another section of this report, it should be noted that the final choice of model fell on the AdaBoost Classifier, with the following parameters: boosting algorithm: SAMME; base estimator: Logistic Regression, learning rate = 0.1 and number of estimators = 100.

# Results

The **Results** section organizes the findings [4]. This section is the meat of a report, the most important part of a study. All other sections serve subordinate roles, either preparing the reader for the **Results** or providing supplemental information to augment the findings. **Results** are general statements that present the research's key results (data) without interpreting their meaning. The author should not include the raw data but present them as text, illustrations, and tables. All these three forms may be used, but the same data should not be repeated in more than one form. The results of statistical analyses should also be stated in this section [4]. Consider the following guidelines when writing the **Results** section [4]:

1. It is not necessary to include all the collected data during the research. This isn’t a diary. Instead, select and emphasize only important and relevant data that will answer the question or solve the problem raised in the Introduction section.
2. Do not include information properly belonging to other sections of the paper such as Materials and Methods, or Discussions (if Results and Discussions are separated).
3. Prevent repeating the legends for figures or the titles of tables in the text.
4. Explain in the text only those illustrations and tables whose significance is not obvious to the reader. Important features that are readily apparent from the illustrations and tables should be highlighted in the text. Do not repeat the data presented in the illustrations and tables.
5. Be sure that the text, illustrations, and tables are consistent with one another. In addition, make sure that all numerical values in every table agree with the figures or data presented.
6. Analyze your data by statistical methods, if appropriate.

Be honest. Do not omit data that do not support your hypothesis and conclusion or do not answer the research question.

To conclude, the results section of a report has two key features: an overall description of the study's major findings; and the data should be presented clearly and concisely [3].

# discussion

The Discussion section attaches the findings to other existing scientific papers to form new ideas [4]. The Discussion section of a scientific article reiterates the main findings but in the context of furthering knowledge or impacting on teaching practice, or future research. In other words, the **Discussion** takes and interprets the findings reported in the Results section, evaluates their significance, and examines the implications. This is probably the most challenging to write among all sections in a research article and will demonstrate how well the author understands the results. Nevertheless, it does not mean that the Discussion should be long, especially if there is little to discuss. [4]

# Conclusion

The conclusion presents the outcome of the work. In it, you should interpret the findings at a higher level of abstraction than the one you did in the ***Discussion***. Moreover, it would help if you related these findings to the motivation stated in the ***Introduction*** [2]. Finally, this section is where the author restates the contribution of the research, with a particular emphasis on what it allows others to do; and proposes new research directions to prevent duplication of effort or encourage collaboration[4].

# References

Every research project usually relies in part upon the work of other scientific works. Therefore, any time an author cites external materials, he/she must identify his/her sources in the form of systematic references. The importance of the References section in a report is not only for giving credit to the ideas and work of other scientists but also to provide the readers with access to these sources [4]. [You should use a reference management system such as Endnote, Mendeley, Citavi or Zotero.]

[1] Robert A Day. How to write and publish scientific reports, 1998.

[2] Jean-Luc Doumont, Laura Grossenbacher, Christina Matta, and Jorge Cham. English communication for scientists. 2014.

[3] George M Hall. How to write a report. John Wiley & Sons, 2012.

[4] Parlindungan Pardede. Scientific articles structure. In Scientific Writing Workshop, volume 16, 2012.

# Appendix (optional)

*[An* ***annex*** *can stand alone. For example, if you are attaching additional documents to the end of your report that would make complete sense and provide important information even outside the context of your report, you can categorize them as annexes.*

*An* ***appendix*** *tends to be more closely connected than an annex to the main body of the report. An appendix would not be as informative or valuable outside the context of your report. While an appendix enhances or expands upon your research report by adding details like illustrations or case studies, it is never presented to readers by itself.*

*Retrieved from* [*https://www.enago.com/academy/annex-vs-appendix-do-you-know-the-difference/*](https://www.enago.com/academy/annex-vs-appendix-do-you-know-the-difference/)*]*

# Annexes (optional)

