

**Group Project**

**The Smith Parasite**

**Machine Learning 2022/2023**

**Work group:**

-Beatriz Carmo | Student Number: 20220685

-João Maio | Student Number: 20220696

-Lizaveta Barysionak | Student Number: 20220667

-Marta Antunes | Student Number: 20221094

-Tomás Silva | Student Number: 20221639

# Index

[Introduction 3](#_Toc119358152)

[Exploration 3](#_Toc119358153)

[Feature Selection 4](#_Toc119358154)

[Modelling 4](#_Toc119358155)

[Preprocessing 5](#_Toc119358156)

[Assessment 6](#_Toc119358157)

[Conclusion 6](#_Toc119358158)

[References 7](#_Toc119358159)

[Annexes 8](#_Toc119358160)

# Introduction

This project objective is to build a model that answers the question; “Who are the people more likely to suffer from the Smith Parasite?”.

To answer this question, we started by analyze and cleaning our data.

We build a model that can detect who are more likely to suffer from the Smith Parasite.

# Exploration

**Data exploration**

The most important is to understand the variables mean and if we have missing, duplicated values and put everything equal for example lower case. We also check if all the features we have make sense for the classification problem for example birth year to age and making intervals for ages.

**Feature Engineering**

We decided to create new features that seemed more relevant for this classification. The main objective is trying to have the ‘perfect’ dataset possible.

1ªFeature:

# Feature Selection

Metric features

Considering feature selection of the metric features, we use

Categorical Features

Regarding

# Modelling

We tested several metrics to obtain the best model possible for our model.

1. Logistic Regression

We tested Logistic Regression and the model accuracy is 0.825

Uma imagem com texto

Descrição gerada automaticamente

1. Decision Tree Model

# Preprocessing

# Assessment

# Conclusion

Our final model is

# References

# Annexes