



# Master Thesis in department of Bioinformatics

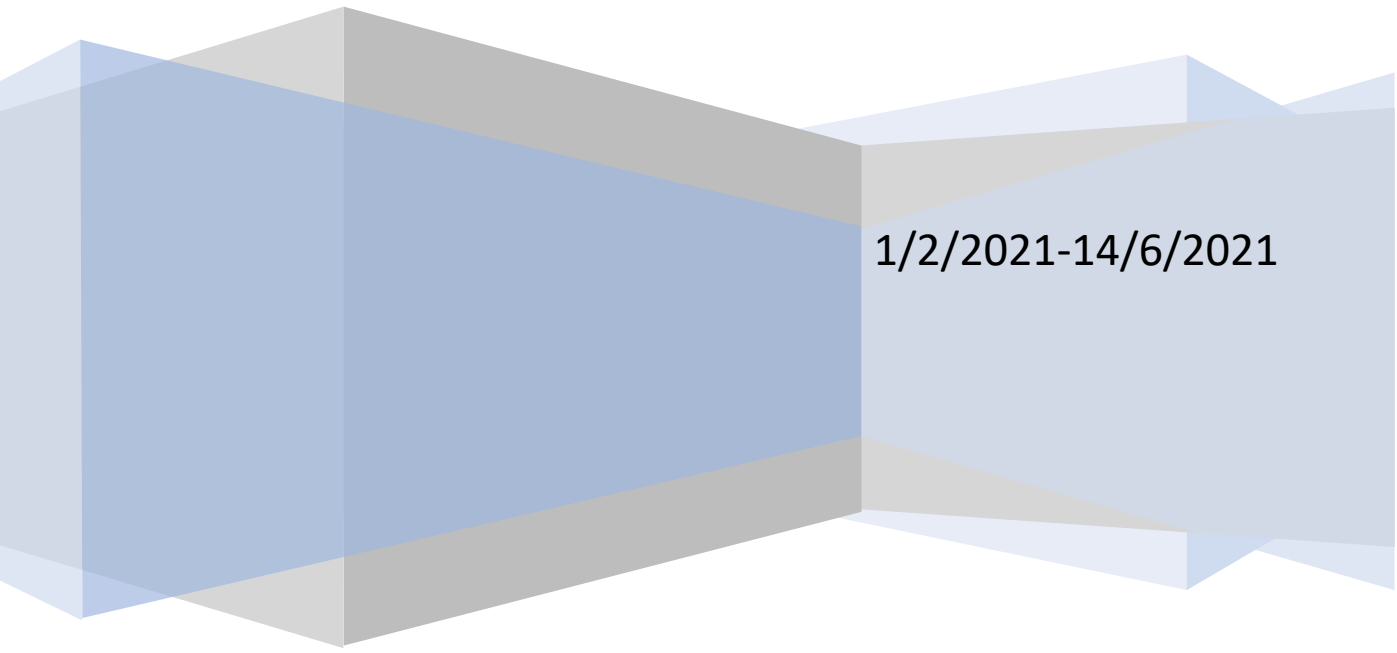
## Missing data in questionnaires - a machine learning study

*Student: Eirini Chatziloudi*

*Student Nr.: 201903689*

*Supervisor: Palle Villesen*

*ECTS: 30 ECTS*

A decorative graphic at the bottom of the page consisting of several overlapping, semi-transparent geometric shapes in shades of blue and grey, creating a modern, abstract design.

1/2/2021-14/6/2021

# Acknowledgments

I specially thank my supervisor Palle Villesen for his help and his motivation. The Bioinformatics Research Center culture is amazing and even in this difficult period with remote meetings and lessons, the teachers tried to be very close to the students with patience and understanding. Finally, I want also to thank Lotte Aas Hindhede from the Danish Blood Donor Study in Aarhus University Hospital, because she was always available when we needed her help.

# CONTENTS

ABSTRACT .....	4
INTRODUCTION .....	4
The Danish Blood Donor Study .....	4
Missing data in questionnaires .....	4
Aim of this project.....	5
Machine Learning.....	5
Random Forest.....	6
Caret Package.....	6
Cross validation .....	7
METHODS & RESULTS .....	7
Data preprocessing .....	7
Data Description.....	7
Data Cleaning .....	9
Questions weighting .....	15
Implementation of Machine Learning .....	16
Handling with missing data .....	16
Implementation of predictive models .....	17
Comparing different methods .....	21
DISCUSSION.....	22
CONCLUSION.....	23
CODE AVAILABILITY .....	23
APPENDIX .....	24
REFERENCES .....	28

# ABSTRACT

**Background:** In this thesis project we study the answers from the questionnaires given by the participants in the Danish Blood Donors Study. We try to find patterns in the questions that are not answered, why they are not answered and by whom. We then try to predict the missing answers based on the answers to the previous questions by using different predictive models.

**Methods:** The data from the questionnaires were given in 3 different tables. We explore one table of them (the DBDS1 table) and we first try to clean the data. After having found some patterns on which questions are not usually answered we implement some machine learning methods in order to predict the missing answers.

**Conclusion:** After our data preprocessing we managed to reduce the rate of missing answers and also to find some simple patterns which predict questions are not going to be answered. Our predictive models managed to give us some first predictions for some questions with an acceptable RMSE, but still they need a lot improvement.

# INTRODUCTION

## The Danish Blood Donor Study

The Danish Blood Donor Study (DBDS) is a research project that aims to benefit both blood donors and Danish patients in the fight against diseases. The Danish Blood Donor Study is a national multicenter study and all the five blood centers in Denmark participate. Approximate more than 110,000 blood donors are currently included in the study. Upon inclusion, a questionnaire is completed and a whole blood sample and a plasma sample are stored. In addition, a plasma sample is stored at each donation, which is why a total of more than 1 million plasma samples have been collected from participants.

From participants, questionnaire data are available in many different areas (e.g. self-reported health (SF-12), smoking, BMI, infections, sleep, stress, allergies, etc.). More than 300,000 responses have been completed. In addition, the study has access to register data via Statistics Denmark, e.g. the Danish National Prescription Registry (Igemiddelregistret) and the Danish National Patient Registry (Landspatientregistret). In relation to blood donation, various blood values are also measured, including levels of hemoglobin and iron. In collaboration with Icelandic colleagues at deCODE Genetics, the first 110,000 registered donors have been genotyped for more than 600,000 gene variants using the Illumina Global Screening Array.

The combination of questionnaire data, registry data, biological measurements and genetics creates a unique research base, with ample opportunity to work with a lot of data in many different ways. [9] [10]

## Missing data in questionnaires

Via questionnaires, the participants have contributed with over 300,000 answers and a number of blood values are measured. A challenge in large population studies is the lack of data: Can patterns be found in which questions are not answered? Who does not answer which questions? How well can one predict the answer to a question based on the answers to the other questions?