Predicting the outcome of a PCR test for COVID-19 using a routine blood exam: A reimplementation

An example of how AI can help identify possible infections in a primary care/triage system

By

Lukas Schießer

Submitted in partial fulfillment of the requirements for the degree of
Bachelor of Science in Cognitive Science
to the
Institute of Cognitive Science at the Osnabrück University
December 6, 2020

Thesis Supervisor:
Johannes Schrumpf M. Sc., Institute of Cognitive Science, Osnabrück University
Thesis Supervisor:
Dr. Tobias Thelen, Institute of Cognitive Science, Osnabrück University

Abstract

Acknowledgements

I want to thank...

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Introduction for this thesis by explaining motivation for this topic and giving a short overview.
Motivation: overwhelmed health care systems, shortages of testing supplies, AI as supplementary technique to help.

Chapter 2

Materials and Methods

2.1 Data

The data used to train the classifiers was provided by Brinati et al. [2]. It was collected between the end of February 2020 and mid of March 2020 from patients admitted to the *IRCSS Ospedale San Raffaele* and consists of 279 individuals who were selected randomly. For each individual, the data set provides their age, gender, results of a routine blood screening, and the result of a PCR test for Sars-CoV-2. A complete overview over the recorded features is provided in A.1. The target variable *Swab* is binary and indicates the result of a PCR-test for Sars-CoV-2 taken by nasopharyngeal swab. A 0 indicates a negative test while a 1 indicates a positive test. The data set is slightly imbalanced towards positive cases with 102 (37%) negative cases and 177 (63%) positive cases.

Since the variable *Gender* was provided as a string, it was transformed into two binary numerical variables called *female* and *male* by one-hot encoding. Further, two values of the variable *Age* were removed, specifically the values 0 and 1. This was sensible seeing that there was no other data recorded from minors under the age of 18 and thus these two values can be presumed to be input errors during the collection process.

Table A.4 provides common statistics for the numerical features of the data set. As you can see in Figure A.1, most of the data is non-normally distributed. Table A.3 shows that most features have missing values. 196 samples have at least one feature missing which amounts to 70 % of the data. Due to the small size of data set it is not feasible to exclude these individuals from the analysis process. It is rather more constructive to use an imputation method that models the missing values based on the observed values in the data set. Therefore, Brinati et al. chose to use *Multivariate Imputation by Chained Equations.*

2.2 Multivariate Imputation by Chained Equations

Multivariate Imputation by Chained Equations or *MICE* for short is an imputation method proposed by Buuren and Groothuis-Oudshoorn [4], it is also known as fully conditional specification (FCS). MICE is a method that imputes missing data by estimating a set of possible values from distributions of observed data. Each variable with missing data x_n is regressed on all other variables $x_1, ..., x_k$ which are restricted to the occurrences with observed data in x_n.

The imputation process is based on the following four main steps [5, 1]: Firstly, all missing values are imputed using a simple imputation method (e.g. mean imputation). These imputations can be thought of as "place holders" used during the first modeling phase. During step 2, the "place holder" imputations for one variable x are set back to missing. In step 3, all observed values from variable x in step 2 are regressed on ther other variables in the imputation model. Since this is the model building phase, this step only uses samples where x has observed values. Therefore, x is the dependent variable and all other variables are independent variables used in the regression model. In step 4, the missing values in x are replaced with imputations (predictions) from the regression model built in step 3. All values of x, the observed and the imputed values, are then used in subsequent regression models of other variables.

Steps 2-4 are repeated for every variable with missing data. After the algorithm is done cycling through all variables, one iteration or "cycle" is completed. Steps 2-4 are repeated for a user-specified number of cycles. Generally, ten to twenty cycles should suffice to stabilize the results of the imputation that is the parameters controlling the imputations should have converged by then. This imputation process is usually repeated m times creating m slightly differently imputed data sets which are then used in the susbequent analysis. According to [3, 1, 5], already a small number of imputed data sets, usually three to ten, is sufficient to provide sensible results during analysis. MICE assumes that the data is missing at random (MAR) that is the probability of data being missing does not depend on the unobserved data but is only dependent (conditional) on the observed data.

> Describe MICE algorithm (keine Bewertung, findet in Discussion statt)
> Short description of implementation in Python using rpy2
> Shortly mention that R implementation is not able to apply MICE model to other data only to data it is "trained"/ "fitted" on (or maybe that's for the 4 Discussion section)

Maybe include PMM?

2.3 Model selection

2.3.1 Random Forest

2.3.2 Logistic regression

2.4 Model training

Describe here k-fold nested cross validation, evaluation metrics used

Here maybe also implementation of MICE in Python using rpy2

Chapter 3

Results

3.1 Results of own implementation

3.2 Comparison with original paper

Chapter 4

Discussion

4.1 Discussion of Results

4.2 Discussion of Methods

4.3 Scenarios for real-world validation

Declaration

I declare that..

Appendix A

Supplementary Tables

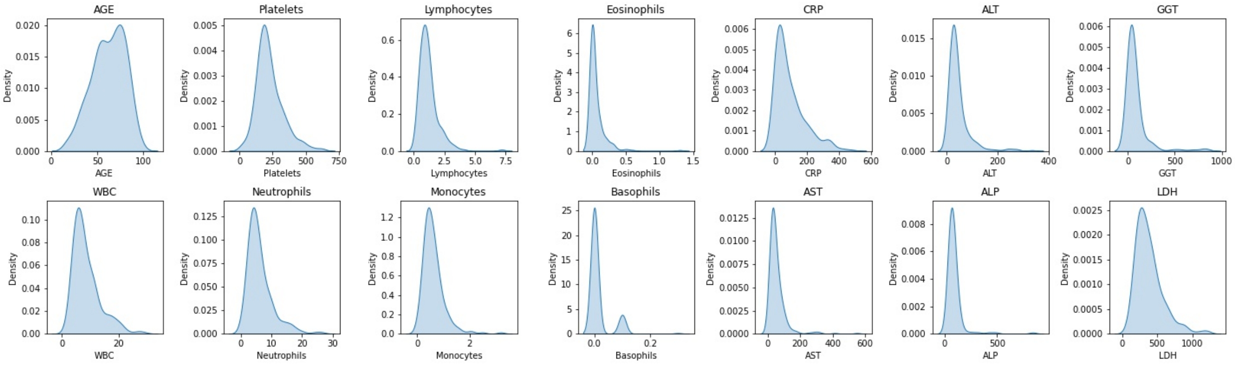

Figure A.1: Kernel density plots of the numerical features of the data set

Feature	Data Type
Gender	Categorical
Age	Numerical (discrete)
WBC (White blood cell count)	Numerical (continuous)
Platelets	Numerical (continuous)
Neutrophils	Numerical (continuous)
Lymphocytes	Numerical (continuous)
Monocytes	Numerical (continuous)
Eosinophils	Numerical (continuous)
Basophils	Numerical (continuous)
CRP (C-reactive protein)	Numerical (continuous)
AST (aspartate aminotransferase)	Numerical (continuous)
ALT (alanine aminotransferase)	Numerical (continuous)
ALP (alkaline phosphatase)	Numerical (continuous)
GGT (gamma glutamyl transferase)	Numerical (continuous)
LDH (lactate dehydrogenase)	Numerical (continuous)
SWAB	Categorical

Table A.1: Overview over all features of the data set

Feature	Unit	Mean	Std	Median
Age	Years	61.33	18.05	64
White Blood Cell Count (WBC)	10^9/L	8.49	4.89	7.10
Platelets	10^9/L	224.91	102.61	204.00
Neutrophils	10^9/L	4.64	4.50	3.90
Lymphocytes	10^9/L	0.88	0.87	0.80
Monocytes	10^9/L	0.45	0.44	0.40
Eosinophils	10^9/L	0.04	0.12	0.00
Basophils	10^9/L	0.01	0.03	0.00
C-reactive protein (CRP)	mg/L	88.93	94.32	53.10
Aspartate Aminotransferase (AST)	U/L	53.81	57.59	36.00
Alanine Aminotransferase (ALT)	U/L	42.82	45.43	30.00
Alkaline Phosphatase (ALP)	U/L	42.21	75.71	68.00
Gamma Glutamyl Transferase (GGT)	U/L	40.20	101.29	0.00
Lactate dehydrogenase (LDH)	U/L	264.54	238.53	254.00

Table A.2: Descriptive statistics for numerical features in data set (including missing values as in [2])

Feature	Number of NaN (in %)
Gender	0 (0 %)
Age	2 (0.72 %)
WBC (White blood cell count)	2 (0.72 %)
Platelets	2 (0.72 %)
Neutrophils	70 (25.09 %)
Lymphocytes	71 (25.45 %)
Monocytes	70 (25.09 %)
Eosinophils	70 (25.09 %)
Basophils	71 (25.45 %)
CRP (C-reactive protein)	6 (2.15 %)
AST (aspartate aminotransferase)	2 (0.72 %)
ALT (alanine aminotransferase)	13 (4.66 %)
ALP (alkaline phosphatase)	148 (53.05 %)
GGT (gamma glutamyl transferase)	143 (51.25 %)
LDH (lactate dehydrogenase)	85 (30.47 %)
SWAB	0 (0 %)

Table A.3: Number of missing values and their proportion the total number of data points

Feature	Unit	Mean	Std	Median
Age	Years	61.78	17.81	64
White Blood Cell Count (WBC)	10^9/L	8.55	4.86	7.10
Platelets	10^9/L	226.5	101.2	205.00
Neutrophils	10^9/L	6.20	4.17	5.10
Lymphocytes	10^9/L	1.19	0.80	1.00
Monocytes	10^9/L	0.61	0.41	0.50
Eosinophils	10^9/L	0.06	0.13	0.00
Basophils	10^9/L	0.01	0.04	0.00
C-reactive protein (CRP)	mg/L	90.89	94.42	54.20
Aspartate Aminotransferase (AST)	U/L	54.20	57.61	36.00
Alanine Aminotransferase (ALT)	U/L	44.92	45.50	31.00
Alkaline Phosphatase (ALP)	U/L	89.89	89.09	71.00
Gamma Glutamyl Transferase (GGT)	U/L	82.48	132.70	41.00
Lactate dehydrogenase (LDH)	U/L	380.45	193.98	328.00

Table A.4: Descriptive statistics for numerical features in data set (excluding missing values)

Appendix B

Second appendix

Bibliography

[1] M. J. Azur et al. “Multiple imputation by chained equations: what is it and how does it work?” In: *Int J Methods Psychiatr Res* 20.1 (2011), pp. 40–9. ISSN: 1557-0657 (Electronic) 1049-8931 (Linking). DOI: 10.1002/mpr.329. URL: https://www.ncbi.nlm.nih.gov/pubmed/21499542,https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/pdf/MPR-20-40.pdf.

[2] D. Brinati et al. “Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study”. In: *J Med Syst* 44.8 (2020), p. 135. ISSN: 1573-689X (Electronic) 0148-5598 (Linking). DOI: 10.1007/s10916-020-01597-4. URL: https://www.ncbi.nlm.nih.gov/pubmed/32607737,https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7326624/pdf/10916_2020_Article_1597.pdf,https://zenodo.org/record/3886927#.X3xhPO1CRPY.

[3] Stef van Buuren. *Flexible Imputation of Missing Data, Second Edition.* 2nd ed. Boca Raton, Fla: CRC Press, 2018. ISBN: 978-0-429-96034-5. URL: https://stefvanbuuren.name/fimd/.

[4] Stef van Buuren and Karin Groothuis-Oudshoorn. “mice: Multivariate Imputation by Chained Equations inR”. In: *Journal of Statistical Software* 45.3 (2011). ISSN: 1548-7660. DOI: 10.18637/jss.v045.i03.

[5] I. R. White, P. Royston, and A. M. Wood. “Multiple imputation using chained equations: Issues and guidance for practice”. In: *Stat Med* 30.4 (2011), pp. 377–99. ISSN: 1097-0258 (Electronic) 0277-6715 (Linking). DOI: 10.1002/sim.4067. URL: https://www.ncbi.nlm.nih.gov/pubmed/21225900,https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4067.