UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Machine learning for predicting mortality and morbidity after Traumatic Brain Injury (TBI)

Diploma Thesis

Theiou Vasileios

Supervisor: Stamoulis Georgios

Month 2022

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Machine learning for predicting mortality and morbidity after Traumatic Brain Injury (TBI)

Diploma Thesis

Theiou Vasileios

Supervisor: Stamoulis Georgios

Month 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Μηχανική μάθηση για πρόβλεψη θνητότητας και νοσηρότητας μετά από Τραυματική Κάκωση Εγκεφάλου (ΤΚΕ)

Διπλωματική Εργασία

Θείου Βασίλειος

Επιβλέπων/πουσα: Σταμούλης Γεώργιος

Μήνας 2022

Approved by the Examination Committee:

Supervisor **Stamoulis Georgios**
Professor, Department of Electrical and Computer Engineering, University of Thessaly

Member **Giannakopoulos Georgios**
Researcher, NCSR Demokritos

Member **Kolombatsos Konstantinos**
Assistant Professor, Department of Informatics and Telecommunications, University of Thessaly

Acknowledgements

First and foremost, I would like to express my gratitude to all the members of the examination committee for offering me the guidance and support the were needed to finish and publish my work. Primarily, I want to thank Mr. George Giannakopoulos as he was always available and willing to help me overcome any difficulties that came up, through guidance, valuable advice and encouragement. It was him that offered me the chance to work on such an important topic and throughout this whole process he transmitted a lot of his knowledge and helped me get experience on how research studies and teams work. It means a lot to me that he offered so much of his time and he played an important role in the progress of the Thesis.

I would also like to express how grateful I am to Mr. George Stamoulis, the supervisor of the Thesis, as he encouraged me into collaborating with Mr. George Giannakopoulos and he offered his knowledge and guidance in order to compolete the study. Additionally, I want to show my appreciation to Mr. Konstantinos Kolombatsos for being a member of the commitee.

Moreover, I would like to show my appreciation to the clinicians from the University Hospital of Heraklion, who were responsible of collecting the data and trusted me for the research. They were always willing to help me with any difficulties on medical terms as it was a new field for me. More specifically, I want to thank Mr. Christos Tsitsipanis, Mr. Aris Salapatas-Gkinis and Mr. Athanasios Theofanopoulos for their contribution to this thesis.

Furthermore, I want to thank all of my friends for the support, the empathy and the amazing moments that we went through this journey. Finally, I would like to thank my family for their love, encouragement and support that they offered throughout my studies.

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

«Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I also declare that the results of the work have not been used to obtain another degree. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism».

The declarant

Theiou Vasileios

Diploma Thesis

Machine learning for predicting mortality and morbidity after Traumatic Brain Injury (TBI)

Theiou Vasileios

Abstract

In our days, machine learning has been extensively used to create predictive models in various fields. A very interesting and important application of machine learning relates to healthcare. There are several studies showing that machines can assist clinicians to make treatment decisions and forecast disease outcomes. In this study, we focus on the setting of Traumatic Brain Injury (TBI). Our goal is to develop simple and largely scalable machine learning models that can accurately predict the capabilities of patients 7 days after hospital admission, in order to support the medical practitioner when deciding specific treatments. To this end, we study the capability of different input features to predict the outcome, validating the usefulness of innovative biomarkers, such as interleukins, as significant predictors. We examine 6 different machine learning models, approaching the problem as a supervised classification problem, aiming to target 3 different capability descriptors (Glasgow Comma Scale, Glasgow Outcome Scale and Karnofsky Performance Scale). To develop simple models for TBI outcome prediction and for examining the underlying effectiveness of the predictors, we conduct a performance comparison of our suggested learning approaches. We also examine the effectiveness of varying feature sets, ranging from demographics, to indicators of clinical severity, secondary insults, CT characteristics and interleukins. The promising first results, reaching an F1 micro score of approximately 80% , indicate that this avenue of machine learning exploitation in the TBI setting can be an important addition to the medical arsenal for decision support.

Keywords:

Machine learning, Prediction, Classification problem, Traumatic Brain Injury(TBI), Karnofsky Performance Status(KPS), Glasgow Outcome Scale(GOS)

Διπλωματική Εργασία

Μηχανική μάθηση για πρόβλεψη θνητότητας και νοσηρότητας μετά από Τραυματική Κάκωση Εγκεφάλου (ΤΚΕ)

Θείου Βασίλειος

Περίληψη

Στις μέρες μας, έχουν δημοσιευθεί πολλές επιτυχημένες προβλέψεις μέσω της χρήσης αλγορίθμων μηχανικής μάθησης σε πληθώρα πεδίων. Μία εξαιρετικά ενδιαφέρουσα χρήση είναι στον τομέα της υγείας. Πολλαπλές έρευνες επιβεβαιώνουν ότι η χρήση υπολογιστών μπορεί να βοηθήσει τους κλινικούς γιατρούς σε αποφάσεις και προβλέψεις αποτελέσματος διαφόρων ασθενειών. Στην παρούσα έρευνα, εμβαθύνουμε στο θέμα τη Τραυματικής Κάκωσης Εγκεφάλου. Στόχος μας, λοιπόν, είναι η ανάπτυξη απλών και σε μεγάλο βαθμό επεκτάσιμων μοντέλων μηχανικής μάθησης, τα οποία θα διαθέτουν τη δυνατότητα ακριβής πρόβλεψης των ικανοτήτων των ασθενών 7 ημέρες μετά την εισαγωγή τους στο νοσοκομείο, έτσι ώστε να υποστηρίξουν τους γιατρούς στον σχεδιασμό συγκεκριμένων θεραπειών. Επίσης, μελετάται η ικανότητα των διαφόρων χαρακτηριστικών στην πρόβλεψη της έκβασης, επικυρώνοντας την χρησιμότητα καινοτόμων βιοδεικτών, όπως οι ιντερλευκίνες. Κατά την προσέγγιση μας, μελετόνται πολλαπλά μοντέλα εξετάζοντας το πρόβλημα ως πρόβλημα ταξινόμησης, θέτοντας 3 μεταβλητές ως στόχους(Κλίμακα κόμματος της Γλασκώβης, Κλίμακα Αποτελεσμάτων Γλασκώβης και Κλίμακα απόδοσης Karnofsky). Για την ανάπτυξη απλών μοντέλων πρόβλεψης διεξήγαμε σύγκριση της απόδοσης 6 διαφορετικών αλγορίθμων, οι οποίοι είναι οι εξής : logistic regression, k nearest neighbors, random forest, decision tree, extratree και catboost. Οι παράμετροι που εισήχθησαν στα μοντέλα περιλάμβαναν δημογραφικά χαρακτηριστικά, δείκτες κλινικής βαρύτητας, δευτερογενείς προσβολές, αξονική τομογραφία χαρακτηριστικά και ιντερλευκίνες. Τα πρώτα υποσχόμενα αποτελέσματα πέτυχαν απόδοση περίπου 80 %, γεγονός το οποίο υποδηλώνει ότι η χρήση της μηχανικής μάθησης σε συνδυασμό με συγκεκριμένες παραμέτρους μπορεί να αποτελέσει σημαντική προσθήκη στο ιατρικό οπλοστάσιο για υποστήριξη αποφάσεων.

Λέξεις-κλειδιά:

Μηχανική μάθηση, Πρόβλεψη, Πρόβλημα Ταξινόμησης, Κλίμακα κόμματος της Γλασκώβης, Κλίμακα Αποτελεσμάτων Γλασκώβης και Κλίμακα απόδοσης Karnofsky

Table of contents

Acknowledgements **ix**

Abstract **xii**

Περίληψη **xiii**

Table of contents **xv**

List of figures **xvii**

List of tables **xix**

Abbreviations **xxi**

1 Introduction **1**

1.1 Motivation 1

1.1.1 Contributions 2

2 Background and Related work **5**

2.1 Background 5

2.1.1 Clinical viewpoint 5

2.1.2 Computer science viewpoint 8

2.2 Related Work 13

3 Proposed Method **15**

3.1 Problem Definition 15

3.2 Our Approach 16

3.2.1 Data Exploration 16

3.2.2 Data preparation 17
3.2.3 Data visualization and data insights 19

4 Experiments 37

5 Conclusion 43
5.1 Summary of work 43
5.2 Future work 45

Bibliography 47

List of figures

3.1 Missing Values table 18

3.2 Correlation between continuous numerical variables 20

3.3 Correlation between discrete variables 21

3.4 GCS categories 24

3.5 GOS categories 25

3.6 KPS categories 25

3.7 Level of recovery based on KPS 26

3.8 Level of recovery based on GCS 26

3.9 GCS and KPS distributions on admission 27

3.10 Hypoxia 28

3.11 Hypotension 28

3.12 Pupils reactivity 29

3.13 IL-6 Distributions 29

3.14 IL-10 Distributions 30

3.15 Interleukins grouped by GCS 31

3.16 Interleukins grouped by KPS 31

3.17 Interleukins values for patients that got worse over time 32

3.18 Rotterdam score with GCS 32

3.19 Rotterdam score with KPS 33

3.20 Marshall score with GCS 33

3.21 Marshall score with KPS 34

3.22 Glc grouped by GCS 34

3.23 Glc grouped by KPS 35

4.1 Select 10 Best features 41

List of tables

4.1 Models Evaluation for GCS 38

4.2 Models Evaluation for KPS 39

4.3 Models Evaluation for GOS 40

Abbreviations

GCS	Glasgow Comma Scale
KPS	Karnofsky performance Scale
GOS	Glasgow Outcome Scale
Glc	Glucose
CT	Computed tomography
IL-6	Interleukin 6
IL-10	Interleukin 10

Chapter 1

Introduction

1.1 Motivation

Traumatic Brain Injury (TBI) constitutes a leading cause of morbidity and mortality all over the world. It impacts millions of individuals and, thus, causes a substantial impact on healthcare and economy, accounting for a significant cost in terms of disability, financial loss and life quality. Hospitals often suffer from overpopulation which subsequently results in scarce medical resources for clinicians to use, especially in a very common incident in low and middle income countries which utilize a reduced number of medical machines. Moreover, the significant investment on medical examinations and prescriptions, could potentially be reduced through the application of machine learning and more specifically predictive models. It is a common experience that hospital admissions can have a significant time and effort cost for all stakeholders. We posit that thanks to the development of technology and the advances in artificial intelligence, prognostic and predictive models could be developed to fine-tune the hospitalization procedures (from admission, to treatment and release). Thus, the purpose of this study is to develop a largely scalable prognostic model based on machine learning algorithms and data form patients with Traumatic Brain Injury, which could give an early prediction of the patient's status after seven days. But first, let us explain what is actually traumatic brain injury, how it is caused and what are its effects on people. Traumatic Brain Injury occurs when a sudden assault damages the brain. It may be caused when an external force or an object penetrates or brakes the skull which can result in a serious injury in the brain. This is called penetrating brain injury. When there is a non penetrating injury with no break in the skull, it is called closed brain injury. Closed brain injury can be caused by a bump, a jolt to the head or a shaking of the brain inside the skull resulting into tearing of the brain tissue and blood vessels. It is a common incident in car accidents, sports, falls etc. There, also, exist three levels of symptoms: mild, moderate and severe. The mild TBI can result in loss of consciousness and dizziness for a few minutes or seconds, while in moderate TBI this may last up to hours, weeks or months along side with confusion. Lastly, the severe TBI is often caused by crushing of the skull and damage the brain. Moreover, the difficulties expand in many aspects of everyday tasks, as TBI patients often face trouble in communicating, concentrating and learning new skills. Brain damage can also affect the emotional status and behavior of patients, coming up with troubles in controlling behavior and personality changes. Finally, severe TBI can cause disabilities in motor skills, hearing and vision of people. These situations can be life-threatening and usually come up with symptoms like repeated vomiting and nausea, slurred speech, weakness in arms and legs, confusion and lack of coordination. All of the mentioned categories of TBI might need medical treatment or, in some cases, hospitalization.

It is now clear that an early prognosis on the level of damage of the patient's brain can help doctors make early clinical treatment plans and try to minimize the implications of TBI. Clinicians that treat patients frequently support therapy decisions on their prognosis evaluation. In patients with traumatic brain injury, the use of computer-based prediction of outcome can help target specific treatment interventions in a personalized manner. The development of prognostic models is useful for several reasons. In terms of clinical relevance first, they could assist both doctors and patients in making treatment decisions. Furthermore, they could be considered as a supportive tool for research purposes and statistical analysis, where scientists need to compare results across diverse groups of patients and injury variations.

1.1.1 Contributions

The contributions of this work are outlined below. Considering the predictors available on admission, our goal is to make use of machine learning in order to develop a simple and largely scalable prognostic model which will predict morbidity and/or unfavorable outcome on the 7th day after admission based on Glasgow comma scale (GCS), the Karnofsky Performance Scale (KPS) or the Glasgow Outcome Scale (GOS). GCS is the impairment of conscious level in response to defined stimuli. There are a number of schemes to stratify the severity of head injury. Any such categorization is arbitrary and will be imperfect. A simple system based only on GCS score is as follows: mild (GCS 14–15), moderate (GCS 9–13), severe (GCS ≤ 8). Similar to GCS but on a more detailed scale which varies from 0 to 100, KPS is an assessment tool for functional impairment and is used to describe patient clinical status after a TBI. On a similar manner, GOS is a global scale for functional outcome that rates patient status into one of the following five categories: Dead, Vegetative State, Severe Disability, Moderate Disability or Good Recovery. In simple terms GCS, KPS and GOS are all considered as indicators of the patient's status and capabilities after TBI. We aim to perform a preliminary study on the predictability of morbidity and/or unfavorable outcome after Traumatic Brain Injury.

Furthermore, we perform statistical analysis on the available data and target to find out which predictors contributed the most in the outcome prediction. Feature importance is one of our main areas of interest since we examine whether in this preliminary approach we can confirm the predictive capacity of a set of select biomarkers and computerized tomography(ct) characteristics which we are confident that play an important role in the outcome and are not yet taken into account by most clinicians. This set includes features based on ct scans like Rotterdam score and Marshall classification, blood substances, like Glc and interleukins, and indicators like Glasgow comma scale and Karnofsky performance status on admission. To be more specific, both Marshall classification and Rotterdam ct score are metrics which are used for the classification of the severity on head injuries based on structural imaging of the brain by computing tomography. There have been some experiments on other studies that point the predictive power of both of these categorizations. The basic difference between these 2 scores is that Rotterdam score is a relatively recently described metric with the purpose to overcome the limitations of Marshall, like the inability to classify patients with multiple type of injuries. In view of Glucose(Glc), it is the main type of sugar in blood and its the major source of energy in the cells of the body. Finally, interleukins(IL-6 and IL-10) are a type of cykotine expressed by leukocytes and other body cells. IL-6 is crucial for immune cell activation and differentiation, as well as for regulation of metabolism, neural development and survival, synaptic plasticity, ion homeostasis and the development and maintenance of various neoplasms. IL-10 is the most important cytokine in suppressing inflammatory responses to all kinds of auto-immune diseases and over-limiting conditions immune responses. It has been shown in few other studies that increased levels of IL-10 and IL-6 are associated with unfavorable outcome, while there seem to be elevated levels of IL-6 few hours after the injury. We hope that the results will help doctors with the treatment interventions. In this paper we provide preliminary results on a performance comparison study, discuss the usefulness of the models and examine the predictive capacity of innovative predictors through feature importance and statistical tests.

Based on the above, the work conducted in this work can be summarized as follows:

1. Research on clinical terms related to our topic of interest.
2. Research on other studies and the predictors that are widely used.
3. Statistical analysis on the provided clinical dataset.
4. Performance comparison on 6 machine learning algorithms for prediction purposes.
5. Evaluation and identification of feature importance related to prediction models.

The remaining text is structured as follows. In Section 2 we briefly overview the related work. Section 3 presents the experiments and algorithms used along with the evaluation metrics and results, while in section 4 we perform some statistical tests and discus the inshights of the data. Lastly, section 5 contains the conclusion, a discussion about the importance of the topic and some future goals.

Chapter 2

Background and Related work

2.1 Background

In this section we are going to introduce some basic terms and concepts of the study on healthcare and computer science fields, facilitating the understanding of both the problem and the proposed method as discussed later in text.

2.1.1 Clinical viewpoint

First and foremost, for a clinical background we need to define traumatic brain injury. TBI is a form of acquired brain injury that affects how the brain works. It occurs when a sudden trauma causes damage to the brain and is a common cause of disability all over the world. There exist 3 indicators of the patient's status after traumatic brain injury.

The first one is Glasgow Outcome Scale or GCS. The Glasgow Coma Scale is a scoring system and prognostic indicator used to describe the level of consciousness in a person that has suffered from TBI [1]. It is calculated by adding the rating of three parameters, which are Eye Opening (E), Verbal Response (V) and Motor Response (M). The GCS values are varying between three and fifteen and can be classified into three categories such as: Severe(GCS <= 8), Moderate(9 < GCS < 12) and Mild(13 < GCS < 15).

Another not commonly used, but more detailed, indicator is the Karnofsky Performance Scale (KPS). The Karnofsky Performance Status is an assessment tool for functional impairment and is used to describe patient clinical status after a TBI. It's score is varying from 0 to 100 and the higher the KPS score, the more capable the patient is to perform activities. A KPS score of 0 indicates that the patient has died.

The last indicator used to define the patient's status is Glasgow Outcome Scale (GOS). The GOS is a scale for patients with brain injuries. It is used to objectively describe the extent of impaired consciousness in all types of acute medical and trauma patients. This scale categorizes patients into five categories: dead, persistent vegetative state, severe disability, moderate disability and low disability[2]. The vegetative state indicates unresponsiveness and a lack of higher mental functions, while severe disability shows that the patient will need help for daily living. Also, patients with moderate disability will not need assistance in performing daily tasks and can even be employed but may require special equipment. Lastly, the low disability category shows light damage with minor neurological and psychological deficits.

We now need to clarify the meaning of some of the features of interest. The first category consists of predictors that come from computer tomography scans. Computer tomography scans combine a series of special X-rays measurements to produce images of the brain. These images are then collected and interpreted by clinicians on a diagnostic manner to search for any abnormal evidence. Since we have the images of the brain, we can then calculate some scores that define the scale of the damage. Thus, Rotterdam score is a classification metric and it is designed to improve prognostic evaluation on patients with severe and moderate traumatic brain injury[3]. It is calculated by adding the values of 4 independent score elements which are presented bellow:

- Basal cisterns
  - 0: normal
  - 1: sompressed
  - 2:absent
- Midline shift
  - 0: no shift or a shift of less than 5 mm
  - 1: shift larger than 5 mm
- Epidural mass lesion
  - 0:present
  - 1:absent

- Intraventricular blood or traumatic SAH
  - 0:present
  - 1:absent

Another useful metric is Marshall Classification. Similarly to Rotterdam score, Marshall classification is a CT scan derived metric used to predict outcome of patients with traumatic brain injury[4]. It is calculated through the scale bellow:

- 1: No visible pathology seen on CT scan
- 2: Cisterns are present with midline shift 0-5mm and/or lesion densities present, no high or misxed density lesion > 25cc
- 3: Cisterns compressed or absent with midline shift 0-5 mm, no high or misxed density lesion > 25cc
- 5: Any lesion surgical evacuated
- 6: High or mixed density lesion > 25cc, not surgical evacuated

Moreover, we are interested in blood substances that might be helpful predictors in the outcome of patients with traumatic brain injury. These substances include K, Na etc but our main focus is on the levels of glucose. Glucose is the simplest type of carbohydrate and along with fat and protein is one of the primary energy sources of the body. Glucose levels have seen to be increasing during traumas. Hyperglycemia, which is the increased levels of sugar in the blood, can be harmful to the injured brain as it compromises microcirculatory blood flow, increases blood-brain barrier permeability, and promotes inflammation.

Last but not least, we make use of some innovative biomarkers, called interleukins. Interleukins are a type of cykotines that are expressed by white blood cells(leukocytes) and other body cells. Cykotines are molecules that allow cells to talk to each other. In this study we are interested in IL-6 and IL-10. Hence, interleukin-6 is an important mediator of fever and of the acute phase response. It is produced in response to infection and tissue injuries and contributes to host defense through the stimulation of acute phase responses, hematopoiesis, and immune reactions. Interleukin-10 is an anti-inflammatory cykotine and its role is to limit host immune response to pathogens and thus prevent damage to the host and maintain normal tissue homeostasis.

2.1.2 Computer science viewpoint

First of all, we need make clear what machine learning is and how it can be applied to healthcare. Machine Learning is a field of artificial intelligence which has the ability to imitate human intelligent. More specifically, it is based on the idea that machines can learn from the given data, retrieve patters and make predictions. Also, it contains methods and algorithms which help data scientists analyze the data and build models that help discover information and patterns. In our study, we have to deal with a classification problem which refers to a predictive modeling problem where a class label is predicted for a given example of input data. The machine learning models need to be able to recognize objects and separate them into certain categories. These categorization is based on a so called target variable, which defines the categories. Since we have a target variable, we have a labeled dataset and thus we use supervised machine learning algorithms in order to map a given input to a certain output.

For the analysis of the data we performed some statistical tests for finding correlation between features. We introduce these test briefly below.

Firstly, for the correlation of continuous predictors we used Pearson's correlation which is a measure of the strength of a linear relationship between 2 sets of data. It has values between -1 and 1, where -1 indicates negative correlation, 1 positive and 0 no correlation. Positive values mean that if the values of one variable increase then there is an increase on the value of the other. A negative correlation implies that for the increase of one variable, we would expect a fall on the values of the second variable. No correlation means that there is no particular association of the values of the 2 variables. The formula for calculating the Pearson's correlation is:

$$r = \frac{\sum((x_i - \overline{x})(y_i - \overline{y}))}{\sqrt{\sum((x_i - \overline{x})^2(y_i - \overline{y})^2)}}$$

, where:

1. x_i = values of the x variable of the sample
2. $\overline{x}$ = mean of the values of the x variable
3. y_i = values of the y variable of the sample
4. $\overline{y}$ = mean of the values of the y variable

Secondly, ANOVA is a statistical test used to find relationship between one categorical independent variable and one quantitive dependent variable. It identifies the statistical difference by calculating if the means of each groups are different from the overall mean of the dependent variable. This test consists of 2 hypothesis. The null hypothesis(H_0) is that there is no difference among the group means, while the alternate hypothesis(H_a) is that at least one of the groups differs significantly from the overall mean[5]. ANOVA uses the F-test for statistical significance, which compares the variance in each group mean from the overall mean. Thus, it return a p-value and if the p-value is less than the threshold of 0.05, then the 2 variables are correlated.

In order to find the association of 2 categorical features, we make use of chi-squared test. This test basically compares the observed results with the expected ones and its purpose is to determine whether a difference between this 2 results is due to chance or if there is a relationship between the 2 variables. Similarly to chi-squared, Crammers V is a measure of correlation of 2 nominal variables. The output range is [0,1], where 0 means no association and 1 full association. Unlike, Pearson correlation there is no negative relationship, either the variables are associated or no. It is based on the Chi-squared statistic[6].

Moreover, we need to perform some pre-processing on the data before it is fed to models. To be more specific, machines cannot understand labels and thus we need to transform them into numbers. The data that consists of labels is called categorical. Thus, the process of converting labels to numbers is called encoding. There exist a lot of encoders, but in our case we make use of one-hot-encoder, which creates a new categorical column and assigns a value of 1 to the feature of each sample that corresponds to its original category. Also, the fact that the values of the features differ significantly between their ranges because they are measured in different units of measure, can affect the performance of the algorithms. The idea is that variables which are measured in different scales do not contribute equally on the model and add bias. Thus, we make use of standard scaler not only for changing the range of values by scaling, but also make the distribution's standard deviation equal to 1. Particularly, standard scaler standardizes a feature by removing the mean and then scaling to the unit variance, which means dividing all values by the standard deviation. This process results into a distribution with standard deviation equal to 1 and mean approximately 0. A sample's standard score is computed as follows: z = (x - μ)/s, where μ is the mean of the sample and s is the standard deviation.

Since we have clarified some basic background knowledge, we can continue with the explanation of the models that are used. The first algorithm is multivariable Logistic Regression. It is a formula used to establish relationships between dependent and more than one independent variables. Logistic regression uses the logistic sigmoid function to transform its output to return a probability value which can then be mapped to two or more discrete classes. The sigmoid function maps values to probabilities and more specifically it maps any real value into another value between 0 and 1. Mathematically, it is defined as

$$h_\theta(x) = \frac{1}{1 + e^- \theta^T x}$$

Then, the algorithm sets a decision boundary which is a threshold value that indicates the category of the output. It, also, contains a cost function and our goal is to minimize this cost function to get the minimum error in our predictions. The cost function is mathematically described as

$$\log(h_\theta(x))$$

if y = 1 and

$$\log(1 - h_\theta(x))$$

if y = 0

In order to minimize the cost function, we use an optimization algorithm called gradient descent. The final outcome is calculated as:

$$y = a_1x_1 + a_2x_2 + ... + a_nx_n$$

where n is the number of features. Our goal is to find the values of $a_1, a_2, ...a_n$.

Another algorithm that we experiment with is K Nearest Neighbors (kNN). Knn is a machine learning algorithm that uses feature similarity to predict the cluster that the new point will fall into. Clusters are groups of data points that are as similar as possible. K is a number used to identify similar neighbors for the new data point. The basic process of the algorithm is that we first choose a number for k, find the distance of the new point to each of the training data and then find the k nearest neighbors to the new data point. Finally, the algorithm counts the number of data points in each category among the k neighbors and then assigns the new data point to the class that contains the most neighbors. The measure of similarity between 2 points is distance. Since can tune the type of distance to our desired type, euclidean distance proved to work best for our experiments. Euclidean distance is defined as the square root of the distance between 2 data points p and q:

$$d\left(p,q\right) = \sqrt{\sum_{i=1}^{n}\left(q_i - p_i\right)^2}$$

Therefore, the algorithm aims to group training data points into groups based on similarity and then assign every new data point to the dominant group based on k nearest neighbors. The more similar each data point is with each other, the closest the distance, while the more different points are gonna have higher distance.

Next, we tried experimenting with tree based methods. Thus, the first algorithm that we are gonna explain is Decision Tree. Decision Tree is a very popular machine learning algorithm, which operates like creating a tree. The algorithm starts by determining the best feature in the dataset and split the data into subsets that contain the values of this best feature. This splitting process is like defining a node for a tree. Afterwards, recursively generate new tree nodes using each subset created by the splitting on the previous step. Then, keep splitting until we have optimized a certain measure. For a classification problem, the best feature is defined by a formula called Gini Index Function: $E = \sum(p_k * (1 - p_k))$, where p_k are the proportion of training instances of class k in a particular prediction node.

In the same manner, Random Forest algorithm consists of a large number of individual and uncorrelated decision trees that operate like an ensemble. Each of these trees role is to make a class prediction and then the class which is the most frequent result becomes the prediction of the model. The process of sampling subsets with replacements is known as bootstrapping. The basic concept of this algorithm is that a prediction of a ”committee” of trees is going to be more accurate than that of any individual tree.

Extremely Randomized Trees Classifier(Extra Trees classifier) is an algorithm that operates similarly to Random Forest, by constructing multiple de-correlated decision trees or random forests during training over the entire dataset. Basically, this algorithm constructs trees over every observation, but with different subsets of features. Thus, randomness is not provided with bootstrap, but instead nodes are split randomly at each node.

The last predictive algorithm that is used is Catboost. It is a recently realised machine learning algorithm, which has been proven to achieve high accuracy in various problems and it is widely used. It requires little computational power and belongs to boosting algorithms. What it does is that it implements symmetric trees. The procedure is as follows:

1. Firstly, calculate residuals for each data point using a model that has been trained on all the other data points at that time. For the purpose of calculating residuals for various data points, we train several models. The residuals for each data point that the model has never seen before are calculated at the conclusion.

2. Train a model by using the residuals of each point as class values and repeat step 1.

As mentioned before, one of our primary goals is to find out the importance of each predictor in the outcome. This is called feature importance and refers to techniques that calculate a score for all the input features for a given model. This score represents the importance of each feature to the prediction of the model. A higher score indicates that the specific feature will have a larger effect on the model that is being used to predict. There exist various techniques to calculate such scores such as Select K-Best algorithm, while some algorithms provide scores automatically, like tree based methods. Select K Best algorithm is provided by the Scikit-learn library and is used for extracting best features of given dataset according to k highest score. This score can be changed by the parameter 'score_func', and be applied both on regression and classification problems. It is a very useful algorithm that can be used on pre processing in order to point the importance capacity of features.

Furthermore, for the evaluation process and in order to estimate the skill of the model to unseen data, we used a special case of cross validation, Leave-one-out cross-validation. Cross validation is a resampling statistical method of evaluating and comparing machine learning algorithms. It divides the data into k folds used for training and testing the algorithm. The parameter k defines the number of folds in which to split the given dataset. Leave-one-out cross-validation is a special case of cross-validation in which the parameter k is set to the number of instances in the data.

Lastly, we have to go through some basic terms on the evaluation metrics that were used. Due to the size of the data we used various metrics to evaluate the performance of the algorithm on each dataset produced by different indicators. These metrics include sensitivity, specificity, f1 _micro and accuracy. Sensitivity is a metric that evaluates the ability of a model to predict true positives of each available category, while specificity is a metric that evaluates the ability to predict true negatives of each available category. F1_micro is also a metric that takes into consideration both the number of prediction errors and the type of errors that our model makes and is calculated as the harmonic mean of precision and recall. Finally, accuracy is defined as the fraction of the correct predictions to the total number of predictions.

2.2 Related Work

In recent years, the use of prognostic models to predict disease outcomes has significantly increased. There have been plenty researches on the prediction of the outcome after traumatic brain injury which vary between populations and models. As a result, 2 widely known models have been developed, namely the Corticosteroid Randomization after Significant Head Injury (CRASH) and the International Mission for Prognosis and Analysis of Clinical Trials in Traumatic Brain Injury (IMPACT) [7]based on large clinical trial datasets and come out with very accurate predictions[8]. These models were trained on large datasets collected from 11 studies and used some common predictors like age, motor score, hypoxia etc. The difference in our study is that we make use of predictors like rotterdam score, interleukins and indicators on admission along side with the other widely used predictors. There are studies that prove that there is a relationship between these variables and traumatic brain injury outcome with statistical tests and pattern observations, however they are yet to be applied on machine learning. For instance, the study [9] indicates that increased levels of interleukins 6 and 8 on admission were associated with unfavorable outcome. Thus, in our study we aim to confirm these findings along side with the help of machine learning prognostic models.

Moreover, there are a few studies on low-middle income countries (LMIC) which point out that differences in environments and healthcare systems can affect the predictions and differ in terms of optimal models[10][11][12]. Thus, there is a need in developing studies and models whose data has been drawn from specific populations. Some of the most used and well performing models in other studies are Logistic Regression, Support Vector Machines, Naive Bayes classifier, Random Forest, Artificial Neural Networks[13][14]. The results vary from study to study with some papers getting better results with Logistic Regression [15][10], while other state that ANNs and Deep Learning seem to be more accurate[11] [16][14]. Unfortunately, at this time our dataset consists of few registrations, so deep learning will not be optimal. Nevertheless, our study has been approved by the bioethics committee and we are currently waiting to get access on large amounts of data to support our findings. Hence, we present the top 6 models that seemed to work better for our case. Rather than that, we are interested in working with new features to prove that they are important and, so, we will use algorithms which are widely used and feature importance is easy to find. Finally, most authors evaluate their models using Area Under Curve(AUC), accuracy, sensitivity, specificity and calibration.

To sum up, in this work we examine different algorithms for prediction to validate previous findings, but also experiment with 3 different target variables (instead of one which is typically used in the related work). We also study the effect of different input features to the predictive ability of the algorithms. In this exploratory aspect, we focus on the innovative interleukins biomarkers, Marshall classification, Rotterdam score and glucose for validating their contribution to an accurate prediction result. There are some references on the predictive power of interleukins, but are not widely used predictors in prognostic models.

Chapter 3

Proposed Method

Now that we have gone through some basic programming and clinical terms and related studies, we can proceed with exploring more insights about the data, like quality and useful information, and discussing the process of transforming data, so that it can be suitable for a machine learning model. This step is crucial, as it can firstly give us a better understanding of the innovative features and secondly increase the accuracy and efficiency of the machine learning model. This chapter is structured in 4 subsections each of them summerizing either valuable information about the features or key procedures performed on the data.

3.1 Problem Definition

In the recent decade, the application of machine learning techniques in healthcare to predict disease outcome has increased significantly. There have been several studies which focus on the prediction of outcome after traumatic brain injury, like Impact and Crash but none of them has been widely used. Many of these studies experiment with various models and predictors in order to achieve the best accuracy. Models and predictors are not the only things which may alter in different studies, but also the data that each study uses which varies from country to country. There are plenty of lifestyle, environmental, diet and cultural differences in any population and therefore different data samples. In addition, as it is mentioned in other studies models that are trained with data samples from developed countries, present worse performance when they make prediction for data samples drawn from low-middle income countries probably due to differences in the healthcare system. Thus, there is a need for the development of studies and models based on particularly local populations. These studies can be more precise and accurate and can be a motive to study how cultural, lifestyle and health system variations can affect patients with traumatic brain injury.

With that being said, our primary and goal in this study is to develop a supportive tool for clinicians to assist them in making treatment decisions. Secondly, we want to research and point out the importance capacity of each variable into predictions, which can help find their contribution in the effects of the trauma. Last but not least, we want to prove that interleukins which were mentioned in other studies but not used with prognostic models, play an important role in the outcome status of patients with TBI.

3.2 Our Approach

Our data has been collected from patients that are hospitalized in the University Hospital of Heraklion and are collected and interpreted by clinicians. The model includes clinical and demographic variables such as the patient's age, sex, Glasgow coma scale/GCS, Karnofksy Performance Scale etc. Also, we include predictors based on findings from CT scans and some substances contained in the blood tests. Last but not least, our study utilizes and evaluates the usefulness of innovative biomarkers (like interleukins and more), which as far as we know have not been included in other studies and we are confident that will help us improve the accuracy of the predictions and maybe lead to improving the performance of other more widely used models, like Impact or Crash. After exploring the data, deleting some features due to missing values and handling categorical variables, the model ends up consisting of 39 features. All the patient's data that was obtained is kept anonymous in order to keep up with bioethics.

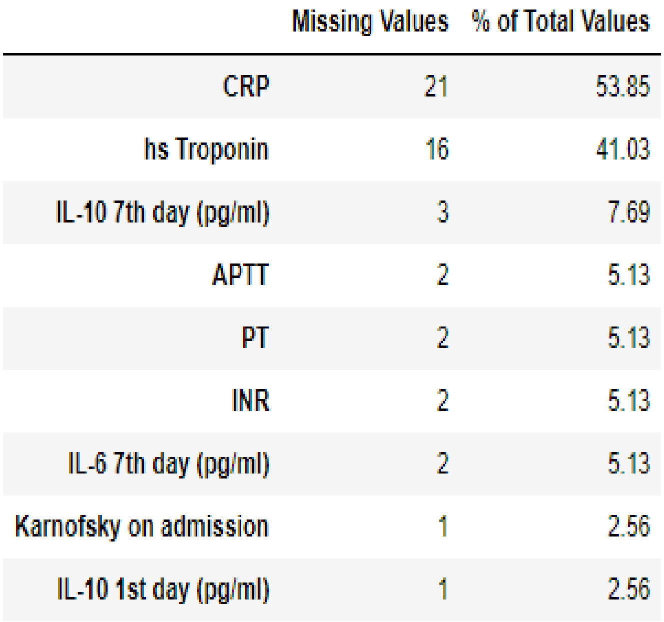
3.2.1 Data Exploration

Due to the fact that we had to deal with healthcare data, the retrieving of the data process was a lot time consuming. Thus, our dataset consists of 39 registrations at the moments, which will increase with time and when the study is approved by the bioethics committee we will have access to much more data, in order to support our findings. By a first glance at the given data we, also, observe that not all the registry data play a part in the model's prediction. Therefore, we needed to explore each variable's definition to choose the ones that are contributing the most. So, all the variables that contained no predictive power(e.g. Date of birth, surgery, reasons of entry, other major injury, type of damage). The dataset contained 3 features that could show the status of the patient and, thus, be considered as target variables. These variables are Glasgow Comma Scale(GCS), Karnofsky Performance Scale(KPS) and Glasgow Outcome Scale(GOS). Each of them can be used as indicators to split data into categories with GCS producing 3 categories(Mild, Moderate, Severe), KPS 10 categories and GOS 5. Each experiment uses one of them as target, while others are dropped due to high correlation with the output. Therefore, when using GCS to define categories, we get that 70% of the patients suffered mild TBI, 23% moderate and 7% severe, which means that we have to deal with highly imbalanced dataset. On the other hand, KPS is made up with 10 categories varying between 0-100 giving as more details about the patient's status, but more sparse data. Lastly, the use of GOS gives us the opportunity to get more detailed outcomes about the patient's status than the GCS and a more balanced dataset.

3.2.2 Data preparation

Our first task before proceeding to the models is to perform data cleaning. Data cleaning is the process of ensuring that our data is correct and useable in order to be fed to the models by identifying any errors or missing data by deleting or correcting it. Firstly, we need to check the data types of each predictor and distinguish them as numerical or non numerical. Due to typos most features of the data were recognized as object type, which we had to convert to a certain type(integer, category or float) in order to continue our exploration. The dataset finally consists of 25 numerical and 12 non numerical features excluding the 3 target variables. For the missing values we firstly corrected the typo mistakes so that the computer realizes that these are nan values and afterwards we set a threshold in each predictor, so those that include missing values over 40% were dropped. Table 3.1 shows the missing values and the percentage of them comparing to the length of the preditor. Missing data can seriously impact the models performance. Hence, predictors with missing values that consist for less than the above threshold, were handled using imputation techniques and were replaced by the mean values. Therefore, CRP and hs Troponin were dropped from the dataset.

Now, after completing the data preparation we can proceed to data preprocessing, which involves transformation of raw data into understandable format. This procedure includes transformations like encoding categorical variables, normalisation, standardisation,feature extraction etc. Most of the predictors included are categorical type and thus need to be en-



	Missing Values	% of Total Values
CRP	21	53.85
hs Troponin	16	41.03
IL-10 7th day (pg/ml)	3	7.69
APTT	2	5.13
PT	2	5.13
INR	2	5.13
IL-6 7th day (pg/ml)	2	5.13
Karnofsky on admission	1	2.56
IL-10 1st day (pg/ml)	1	2.56

Figure 3.1: Missing Values table

coded in order to be passed through the model. We used 2 approaches for the categorical data based on whether each entry can be part of 1 or more categories. For the first case, we used One hot encoder which creates a new categorical column and assigns a value of 1 to the feature of each sample that corresponds to its original category. As for the second case, we converted each entry to a vector whose length is equal to the number of categories and contains 1 on the categories that it is part of and 0 to others. At last, we converted these vector to new features applying the same logic as one hot encoder.

Lastly, in order to finish the pre-processing, we need to apply scaling on the data. This is an critical step due to the fact that features with higher values range dominate other predictors, while we need them to contribute equally. Hence, we make use of standard scaler, an algorithm that scales data such that the distribution is centered around 0 and has standard deviation of 1.

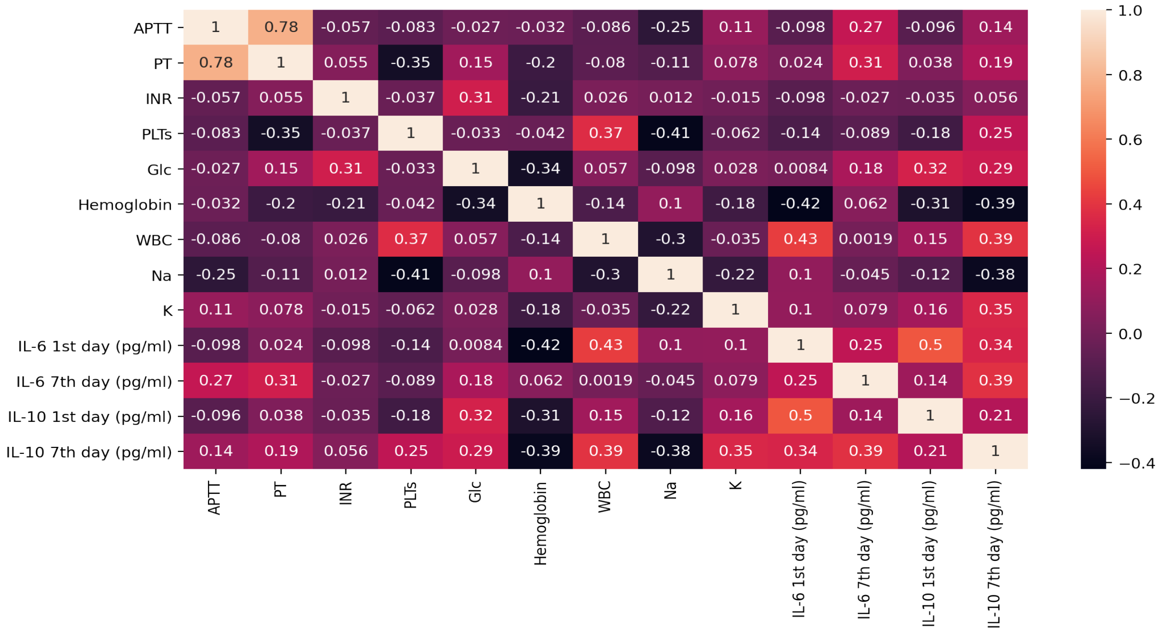
The retrieved data, included the following variables: demographics(age, sex), indicators of clinical severity(GCS on admision, coexistence of major trauma, type of damage, complications, coagulation disorders), pupillary reactivity), secondary insults(hupoxia, hypotension), biochemical variables(APTT, PT, INR, PLTs, Glc, Hemoglobin, WBC, NA, K), CT characteristics(tSAH, Midline shift, Marshall classification, Rotterdam score) and interleukins(IL-6 on 1st day, IL-6 on 7th day, IL-10 on 1st day, IL-10 on 7th day). Also, the target variables can be Glasgow comma scale, Karnofky Performance scale or Glasgow outcome scale.

3.2.3 Data visualization and data insights

Correlation of data points

Correlation is basically a statistical measure that explains how one or more variables are related to each other. It provides us information about the direction of a relationship, the form of the relationship and the degree of strength. It is important in real life problem because if 2 variables have strong correlation then we can predict the value of one with the help of the other variable. Also, 2 highly correlated features have the same effect on the model. so it is suffice to save data for only one and feed it to the model. This can save effort in gathering data and speed up the model. There exist 3 types of correlation relationships: positive, negative and non correlated. Strong positive correlation indicates that as the value of one increases, the value of the other variable increases too. On the other hand, negative correlation indicates the opposite while no correlation means that the change of one variable has no impact on the other. There are different methods to measure correlation based on the type of the variable(continue or categorical). Before we proceed with finding and commenting on correlations we want to make it clear that they are part of the pre-proccesing and indicate a relationship between features, not their predictive power. Thus, it is important to have them in mind but not make conclusions before running the models. Bellow, we present the correlations between the predictors of the dataset.

1. Correlation between continuous numerical variables.



	APTT	PT	INR	PLTs	Glc	Hemoglobin	WBC	Na	K	IL-6 1st day (pg/ml)	IL-6 7th day (pg/ml)	IL-10 1st day (pg/ml)	IL-10 7th day (pg/ml)
APTT	1	0.78	-0.057	-0.083	-0.027	-0.032	-0.086	-0.25	0.11	-0.098	0.27	-0.096	0.14
PT	0.78	1	0.055	-0.35	0.15	-0.2	-0.08	-0.11	0.078	0.024	0.31	0.038	0.19
INR	-0.057	0.055	1	-0.037	0.31	-0.21	0.026	0.012	-0.015	-0.098	-0.027	-0.035	0.056
PLTs	-0.083	-0.35	-0.037	1	-0.033	-0.042	0.37	-0.41	-0.062	-0.14	-0.089	-0.18	0.25
Glc	-0.027	0.15	0.31	-0.033	1	-0.34	0.057	-0.098	0.028	0.0084	0.18	0.32	0.29
Hemoglobin	-0.032	-0.2	-0.21	-0.042	-0.34	1	-0.14	0.1	-0.18	-0.42	0.062	-0.31	-0.39
WBC	-0.086	-0.08	0.026	0.37	0.057	-0.14	1	-0.3	-0.035	0.43	0.0019	0.15	0.39
Na	-0.25	-0.11	0.012	-0.41	-0.098	0.1	-0.3	1	-0.22	0.1	-0.045	-0.12	-0.38
K	0.11	0.078	-0.015	-0.062	0.028	-0.18	-0.035	-0.22	1	0.1	0.079	0.16	0.35
IL-6 1st day (pg/ml)	-0.098	0.024	-0.098	-0.14	0.0084	-0.42	0.43	0.1	0.1	1	0.25	0.5	0.34
IL-6 7th day (pg/ml)	0.27	0.31	-0.027	-0.089	0.18	0.062	0.0019	-0.045	0.079	0.25	1	0.14	0.39
IL-10 1st day (pg/ml)	-0.096	0.038	-0.035	-0.18	0.32	-0.31	0.15	-0.12	0.16	0.5	0.14	1	0.21
IL-10 7th day (pg/ml)	0.14	0.19	0.056	0.25	0.29	-0.39	0.39	-0.38	0.35	0.34	0.39	0.21	1

Figure 3.2: Correlation between continuous numerical variables

In figure 3.2 we calculated correlation between continuous numerical predictors using Pearson correlation. Pearson correlation is used when we want to find out if there is a linear relationship between the variables. Its values range from -1 to 1, where negative values indicate negative linear relationship, while positive ones the opposite. We can see that there is a highly positive correlation between APTT and PT, but not that high to remove one of them. Also, we can see that there is a small positive correlation between both kinds of interleukins and Glc, WBC and each other.

2. Correlation between categorical variables.

   For correlation between categorical variables, we firstly performed a chi squared test to confirm the hypothesis that the features are correlated. Afterwards, in order to find the strength of the relationship we used Crammers V as a measure of association. The output range is [0,1], where 0 means no association and 1 full association. Unlike, Pearson correlation there is no negative relationship, either the variables are associated or no. Thus, figure 3.3 shows the obtained results.

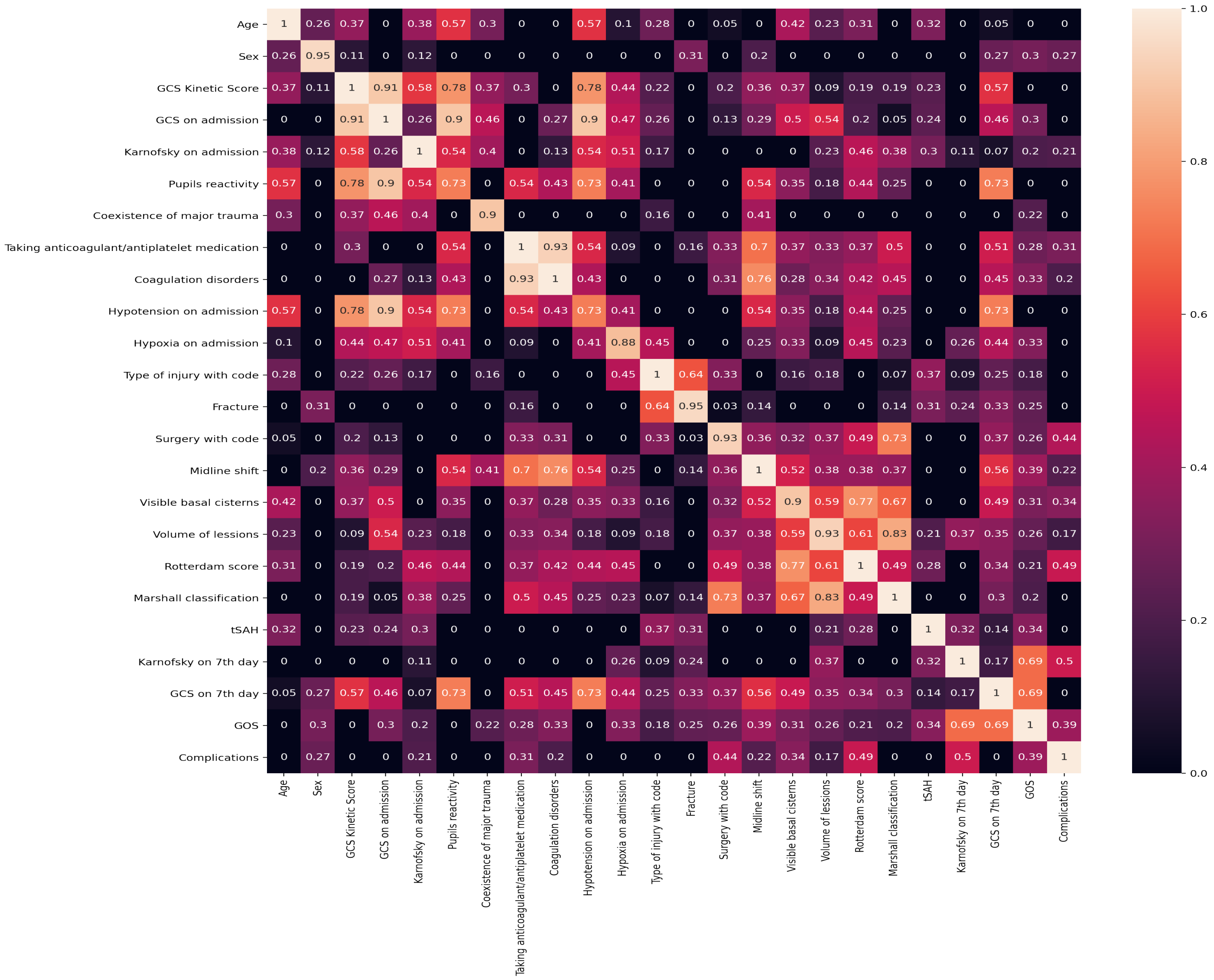


Figure 3.3: Correlation between discrete variables

By taking into consideration the chi square test results and the Crammers V values that are shown in the above image, we can observe that there exist a lot of associations in the dataset. Let us focus on the features of interest and identify their relationship with other variables. First of all, GCS on admission is highly correlated with the pupils reactivity and the hypotension on admission, while it has moderate correlation with hypoxia, Coexistence of major trauma, Visible vasal cicsterns, Volume of lessions and the GCS on the 7th day. However, both tests show that here is no correlation with Marshall and Rotterdam scores. On the other hand, KPS on admission indicates a moderate relationship between these 2 variables along side with hypotension, hypoxia and pupils reactivity. Results on the seventh day, specify that there are not correlations for KPS, but for GCS there are associations with Taking anticoagulant/antiplatelet medication, Rotterdam, Marshall, pupils reactivity, hypotension, hypoxia, Midline shift and Visible basal cisterns. We sum up the correlations on a list so that it can be easier to read with a glance.

- **GCS on admission** Pupils reactivity, Hypotension, Hypoxia, Coexistence of major trauma, Visible vasal cicsterns, Volume of lessions, GCS on the 7th day
- **KPS on admission** Marshall Classification, Rotterdam Score, Hypotension, Hypoxia, Pupils reactivity
- **Marshall Classification** Karnofsky on admission, Taking anticoagulant/antiplatelet medication, Coagulation disorders, Surgery with code, Midline shift, Visible basal cisterns, Volume of lessions, Rotterdam score
- **Rotterdam Score** Karnofsky on admission, Pupils reactivity, Taking anticoagulant/antiplatelet medication, Coagulation disorders, Hypotension, Hypoxia, Surgery with code, Midline shift, Visible basal cisterns, Volume of lessions, Marshall classification, GCS on 7th day
- **GCS on the 7th day** Taking anticoagulant/antiplatelet medication, Rotterdam, pupils reactivity, hypotension, hypoxia, Midline shift and Visible basal cisterns
- **KPS on the seventh day** No correlations

In conclusion, Pupils reactivity seems to play an import on both indicators on admission and on GCS on the 7th day, along side with hypoxia and hypotension. Also, CT characteristics appear to be related both on GCS and KPS.

3. Correlation between numerical and categorical data

   Our last category is to mix things up and find correlation between categorical and numerical data. For this task, we used one-way ANOVA to test the hypothesis that the values are correlated. ANOVA stands for Analysis of Variance and it is a statistical test used to analyze the differences between more than 2 groups. One way ANOVA uses one categorical independent variable and one quantitive dependent variable. It identifies the statistical difference by calculating the means of each groups are different from the overall mean of the dependent variable.

   So, the list presented bellow sums up the associations of the features of importance based on the one way ANOVA test:

   - **GCS on admission** INR, IL-6 1st day (pg/ml), IL-10 1st day (pg/ml)
   - **KPS on admission** Glc, Hemoglobin, WBC
   - **Marshall Classification** WBC, IL-6 1st day (pg/ml), IL-10 1st day (pg/ml)
   - **Rotterdam Score** Glc, IL-10 1st day (pg/ml)
   - **GCS on the 7th day** WBC, IL-6 1st day (pg/ml), IL-10 1st day (pg/ml)
   - **KPS on the seventh day** IL-6 1st day (pg/ml)
   - **IL-6 on the 1st day** GCS on admission, pupils reactivity,Coexistence of major trauma, Hypotension on admission, Surgery with code, Visible basal cisterns, Volume of lessions, Marshall classification, tSAH, GCS on 7th day, GOS
   - **IL-6 on the 7th day** Surgery with code
   - **IL-10 on the 1st day** Coagulation disorders, Surgery with code,Volume of lesions, Rotterdam score, Marshall classification, GCS on 7th day, GOS
   - **IL-10 on the 7th day** Surgery with code, GOS , Complications
   - **Glc** Sex, Karnofsky on admission, Taking anticoagulant/antiplatelet medication, Rotterdam score

An interesting observation is that both inteleukins(IL-6, IL-10) on admission are correlated with GCS on admission and on the 7th day, fact which indicates that interleukins affect the value of the indicator. Furthermore, they are correlated with a lot CT scan information like Marshall, Volume of lessions etc. However, it is pointed out that IL-6 on the first day is associated with more predictors than IL-10, which include symptoms like hypotension, pupils reactivity, hypoxia that are shown to affect the target variables as proven in the previous section.

General information of the sample and indicators

By exploring the data we can observe that the most common incidents that led to head injuries are fall from height and car crashes, where 23 out of 39 patients suffered TBI because of fall from their height probably due to faint. Moreover, we can see that men patients tend to be more than women(33/39), while the mean age of the sample population is 62.5. Our data mostly consists of patients with mild TBI symptoms, as the majority presents a GCS of 15. Figures 3.4, 3.5 and 3.6 show the categories of patients that our data contains based on each of the 3 target variables.

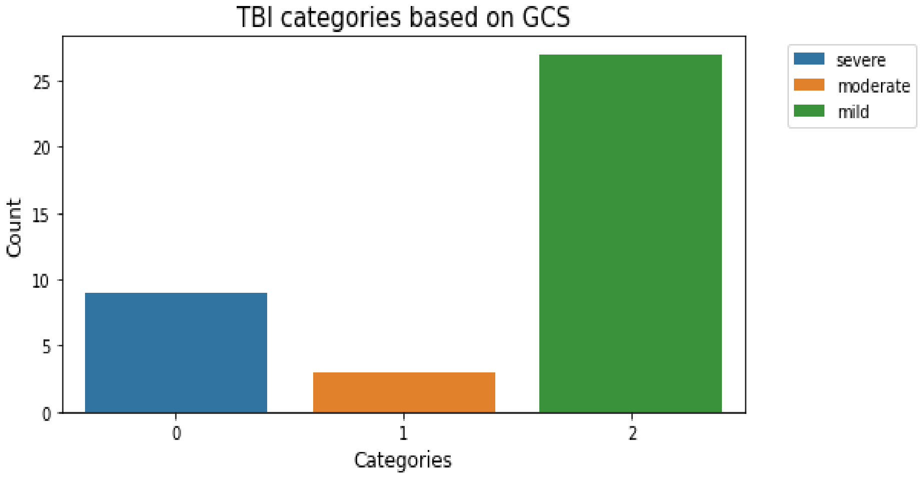


Figure 3.4: GCS categories

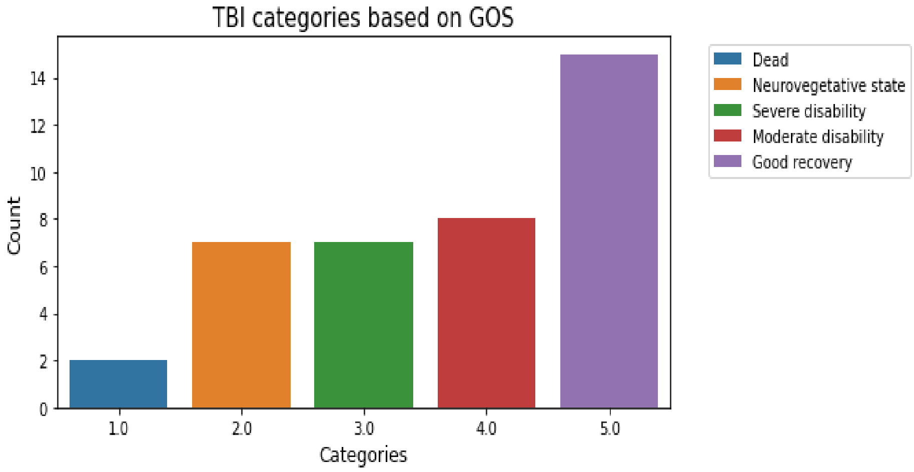


Figure 3.5: GOS categories

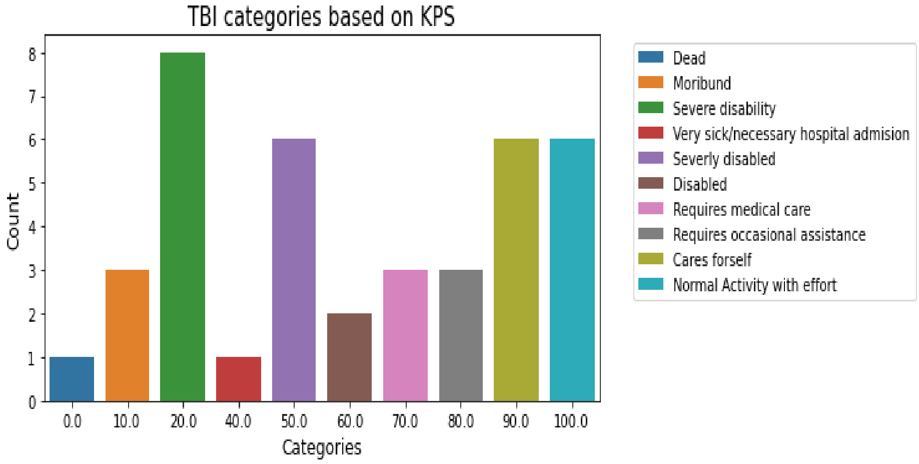


Figure 3.6: KPS categories

An interesting observation would be to visualize the level of recovery of the patients based on GCS and KPS. So, graph 3.7 shows the 2 distributions of the categories based on KPS on the day of the admission(left) and on the seventh day. We can observe that the distribution from right skewed tends to change to left skewed. As shown in the graphs, smaller KPS is more frequent on admission, while on the second graph there are observed higher values. This means that most of our patients were able to recover or improve their status.

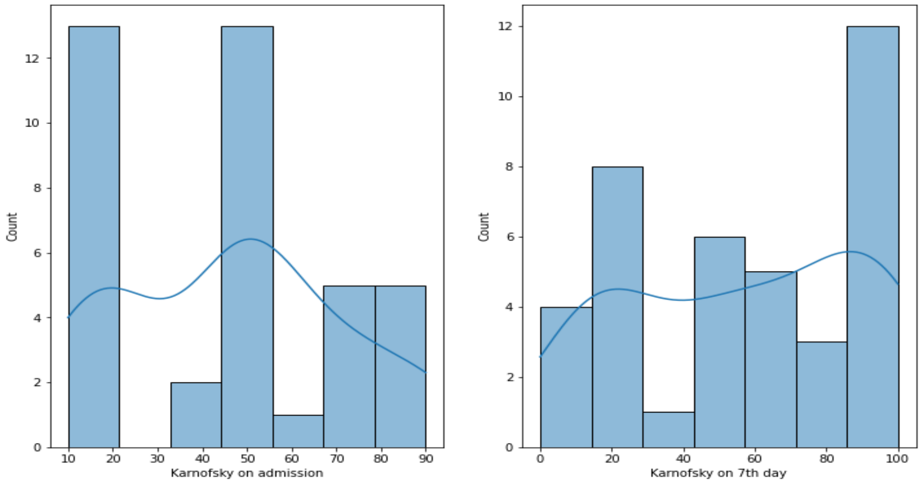


Figure 3.7: Level of recovery based on KPS

In the same manners, figure 3.8 converges to the same conclusion as before. In this chart, we plotted GCS on asmission as x axis and GCS on the seventh day on y axis. The patients are grouped by their GCS score and then we calculate and plot the mean GCS of them on the last day. Apparently, we notice that for every score on admission there is an increase and thus an improvement on the recovery. The black line on the top of each graph indicates the standard deviation of the values, and hence the variation of the data based on each category. High standard deviation is a logical result of the small number of samples on the dataset.

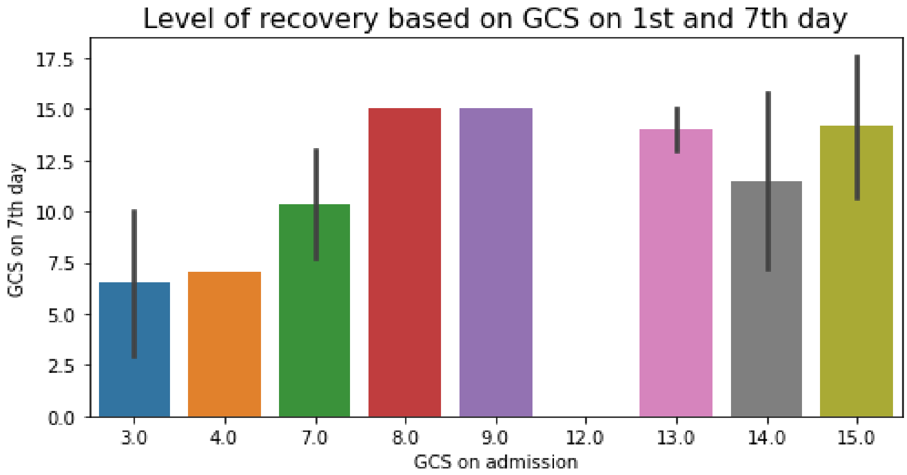


Figure 3.8: Level of recovery based on GCS

Now that we get a general idea about the patients recovery in the seven days interval, we can proceed further investigate the features that we are mostly curious about.

Glasgow Comma Scale and Karnofsky Performance Scale on admission

Firstly, we check the distributions of the status indicators which are presented in figure 3.9. Similar to the values of the seventh day, we can see that for GCS the distribution is left skewed and since the y-axis contains the count of discrete variables, it means that more patients presented higher scores of GCS and thus mild TBI. Also, the registrations of the patients seem to be sparse based on GCS, which might be a problem for the model to identify all classes. On the contrary, KPS distribution is more balanced along values, with most of them being contained in the [40,100] interval.

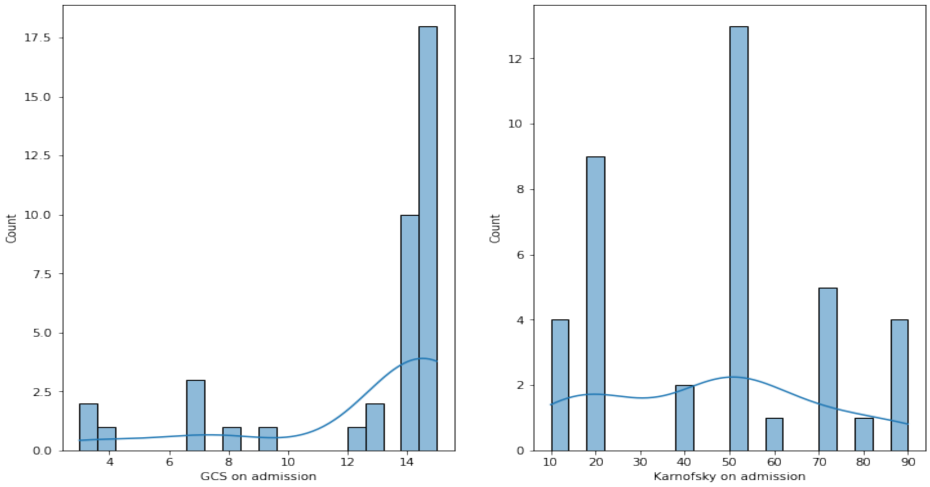


Figure 3.9: GCS and KPS distributions on admission

Furthermore, we can investigate the relationships based on the variables that are correlated to the GCS and KPS on admission from the previous experiments. So, the graphs bellow show the relationship of both indicators with pupils reactivity, hypotension and hypoxia. The three plots indicate that all three anomalies are associated with lower values of both indicators and thus more severe damage to patients.

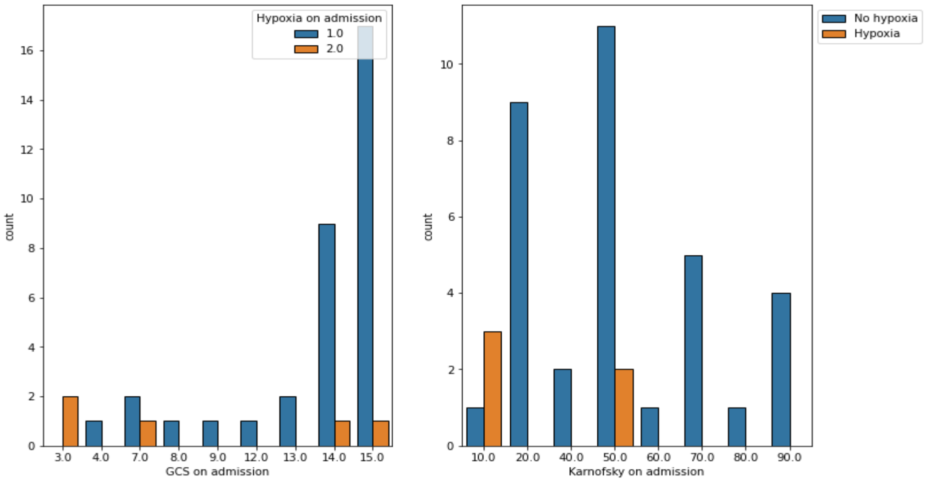


Figure 3.10: Hypoxia

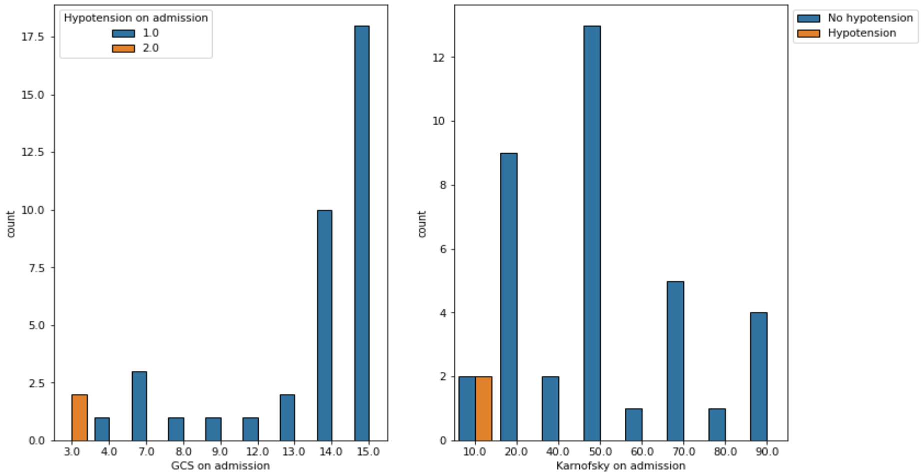


Figure 3.11: Hypotension

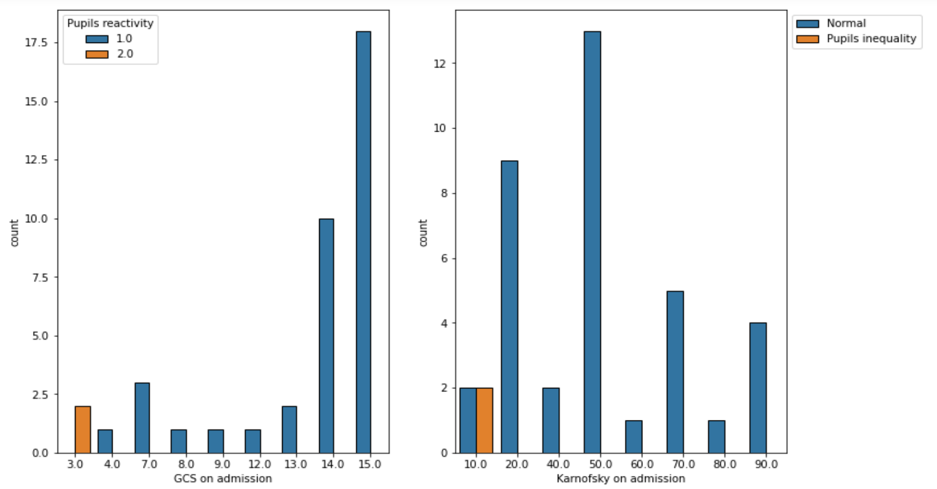


Figure 3.12: Pupils reactivity

Interleukins

As mentioned before, interleukins are very important at the immune cell activation. Let's firstly focus on IL-6. Interleukin-6 is a protein produced by various cells, which helps regulate immune responses and thus makes it a useful marker of the immune system activation. Since our dataset contains values of IL-6 on the time interval, we can plot the distributions as shown in figure 3.13.

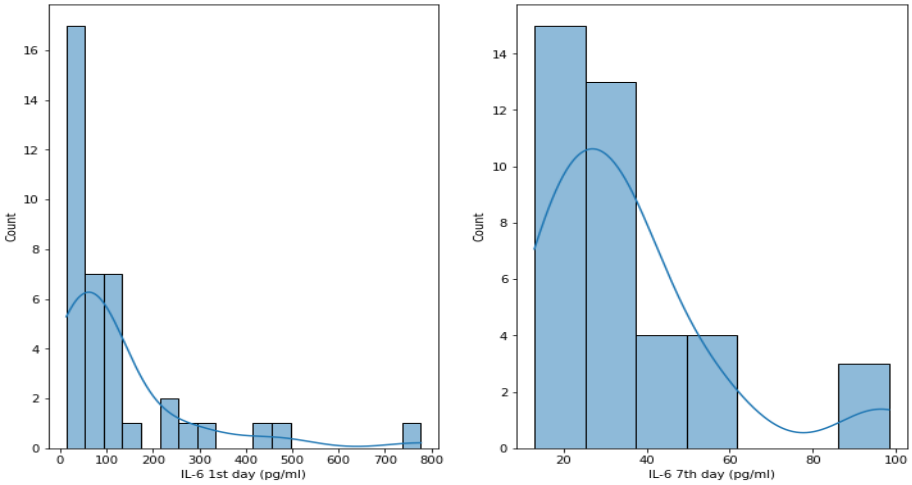


Figure 3.13: IL-6 Distributions

On the other hand, interleukin-10 is an anti-inflammatory cytokine that is essential for stopping autoimmune and inflammatory pathologies. Increased IL-10 levels can impair the host's ability to respond to microbial pathogenesis and inhibit the healing of the tissue damage and hemodynamic abnormalities they cause. Distribution plot for IL-10 is shown in chart 3.14

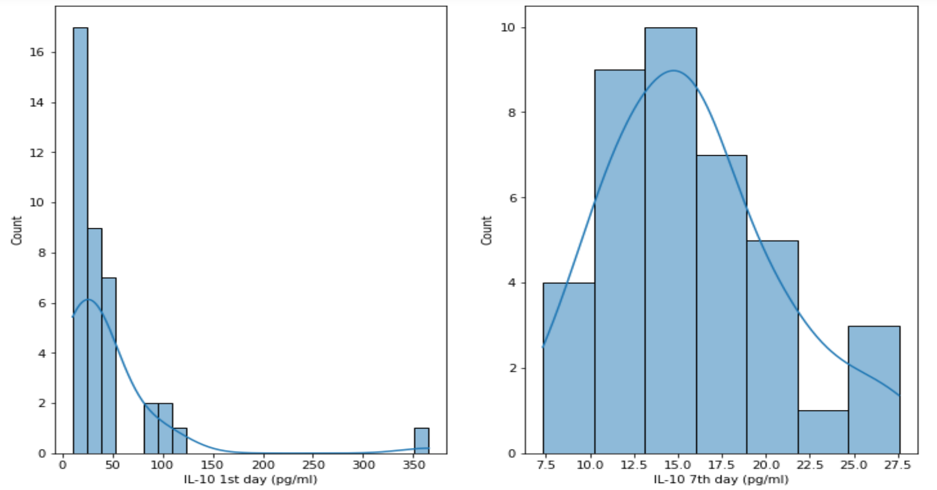


Figure 3.14: IL-10 Distributions

We can note that in both diagrams the values of both biomarkers on the seventh day are much lower than the ones on admission. The peak of the line in each diagram show the mean value. So, for IL-6 we observe means of 122 and 35, while for IL-10 43 and 15. However, a difference is that IL-6 values are mostly clustered on an interval interval of [10,60], while for IL-10 the values are more normally distributed across [7.5, 27.5]. Moreover, taking into consideration the fact that the patients stauts improved an early observation can be that lower levels of interleukins help recover TBI.

Moreover, we can visualize the behaviour of the interleukins on our time interval based on GCS and KPS indicators. That is shown in figures 3.15 and 3.16. So, let's firstly look at GCS which is a less detailed indicator. We can see that more severe categories are associated with higher level of both interelukins. However, categories 2 and 3 which are moderate and mild are very close to each other and similarly do the IL values.Although, we could say the same thing for KPS on admission, on the seventh day there seems to be variations probably due to the fact that this scale is more detailed and we have not many samples. Still, we could point that there exists a fall as the recovery is successful. Similarly, the same phenomenon is noticed with scores on admission.

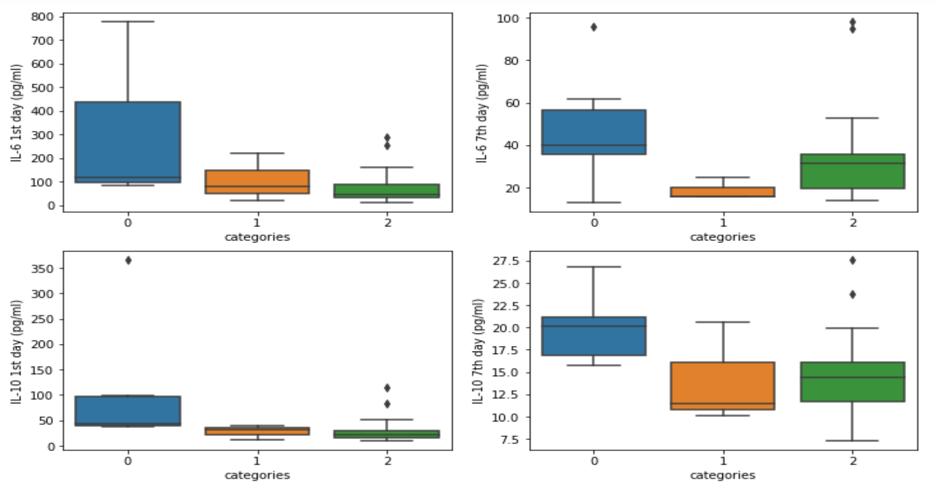


Figure 3.15: Interleukins grouped by GCS

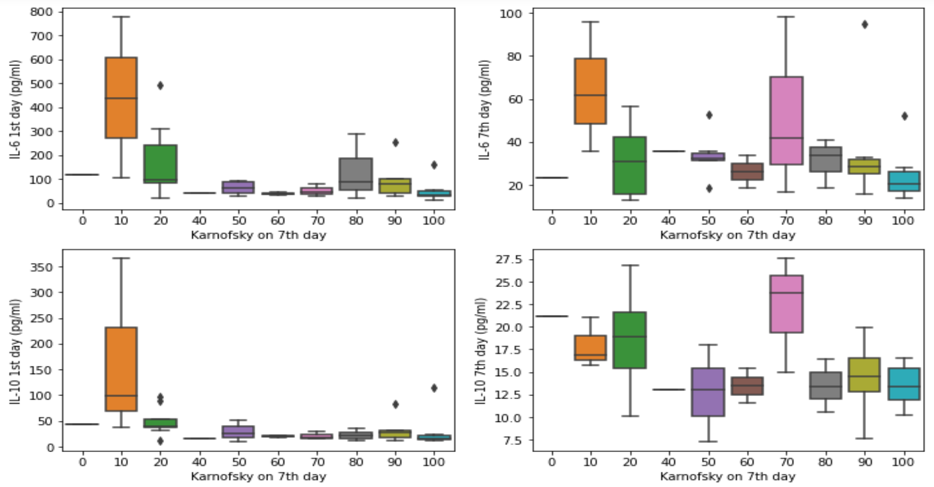
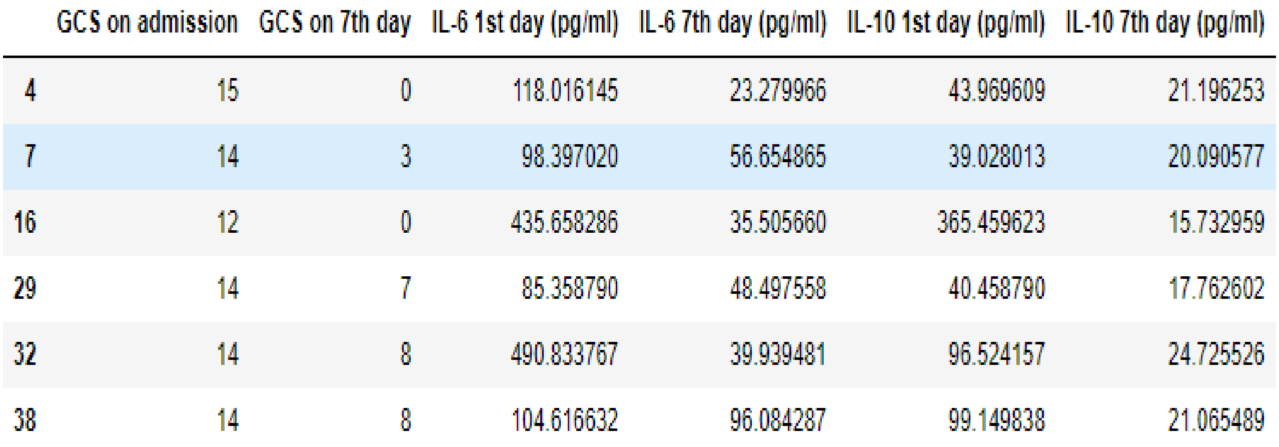


Figure 3.16: Interleukins grouped by KPS

Finally, we can turn our focus on the few patients whose status got worse over the 7 days period to see whether there is any association with the values of interleukins. For the matrix 3.17 we can see that there are 6 patients, whose GCS values were smaller on the 7th day than on admission. For all these cases, the values of IL-6 and IL-10 on admission are higher than the mean values, while for patients that unfavorable outcome, the values were very high. Also, patients who had a very sharp fall on the values of the indicator seem to have big difference in the range of interleukins from admission to 7th day. Hence, these results can be promising on the association of IL-6 and IL-10 with an unfavorable outcome.



	GCS on admission	GCS on 7th day	IL-6 1st day (pg/ml)	IL-6 7th day (pg/ml)	IL-10 1st day (pg/ml)	IL-10 7th day (pg/ml)
4	15	0	118.016145	23.279966	43.969609	21.196253
7	14	3	98.397020	56.654865	39.028013	20.090577
16	12	0	435.658286	35.505660	365.459623	15.732959
29	14	7	85.358790	48.497558	40.458790	17.762602
32	14	8	490.833767	39.939481	96.524157	24.725526
38	14	8	104.616632	96.084287	99.149838	21.065489

Figure 3.17: Interleukins values for patients that got worse over time

Rotterdam and Marshall

Both rotterdam score and marshall classification are metrics that can be calculated through computing tomography, which seem to have predictive power on patients with TBI. Since both of these metrics are discrete, we can visualize the count of patients in each category based again on KPS and GCS. From the figures bellow, we can detect again that there is a trend of higher severe brain damages to be associated with higher both rotterdam and marshall classificaiton scores. Apparently, there are also some patients with mild TBI that have high scores. Probably, more data will help us get more insights on this.

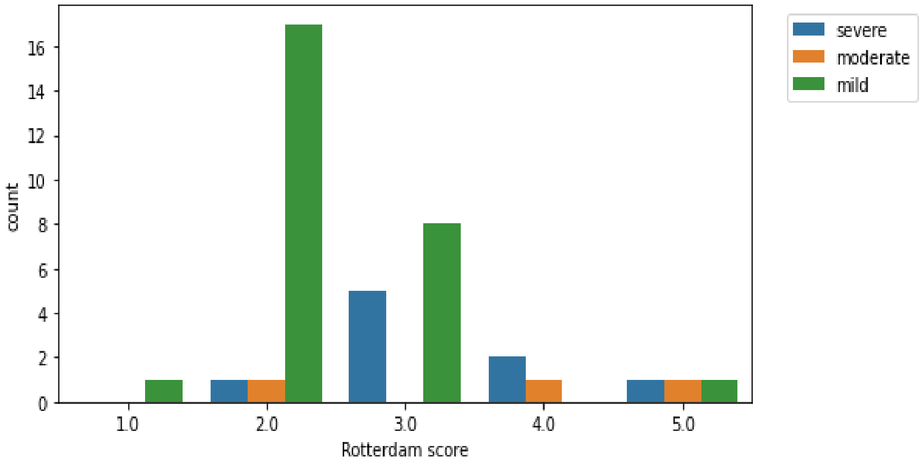


Figure 3.18: Rotterdam score with GCS

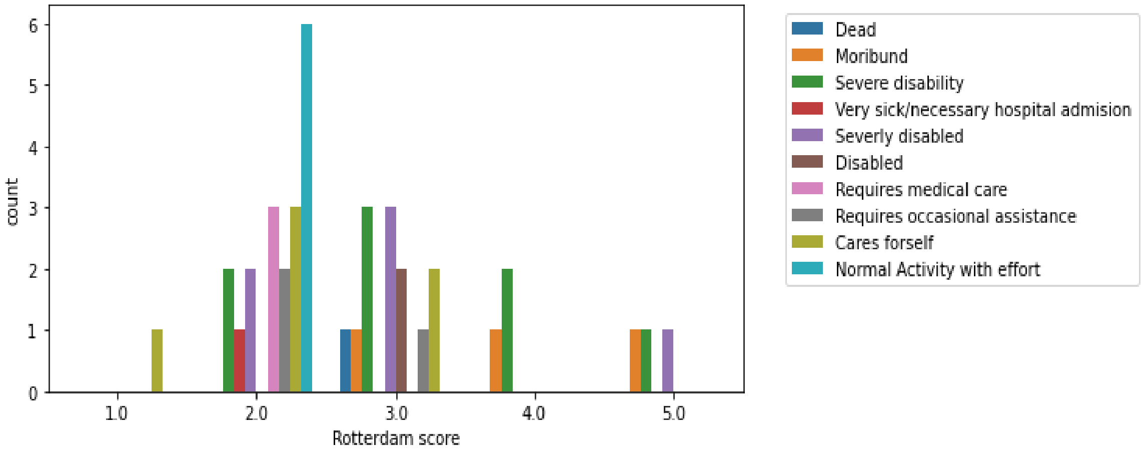


Figure 3.19: Rotterdam score with KPS

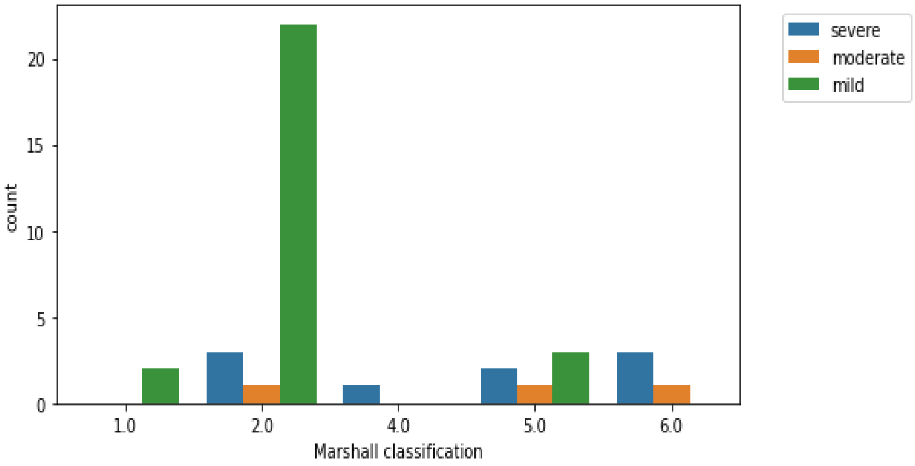


Figure 3.20: Marshall score with GCS

Glucose(GLC)

The last feature that we want to investigate is glc. Glucose is the main type of sugar in the blood and is the main source of energy to our body. We can plot once more the patients status on the 7 day interval along side with the categories in order to check for any useful information. The charts presented indicate that there is a slight fall for higher values of indicators, however we cannot make a statement on that, so we will find more information from the machine learning models.

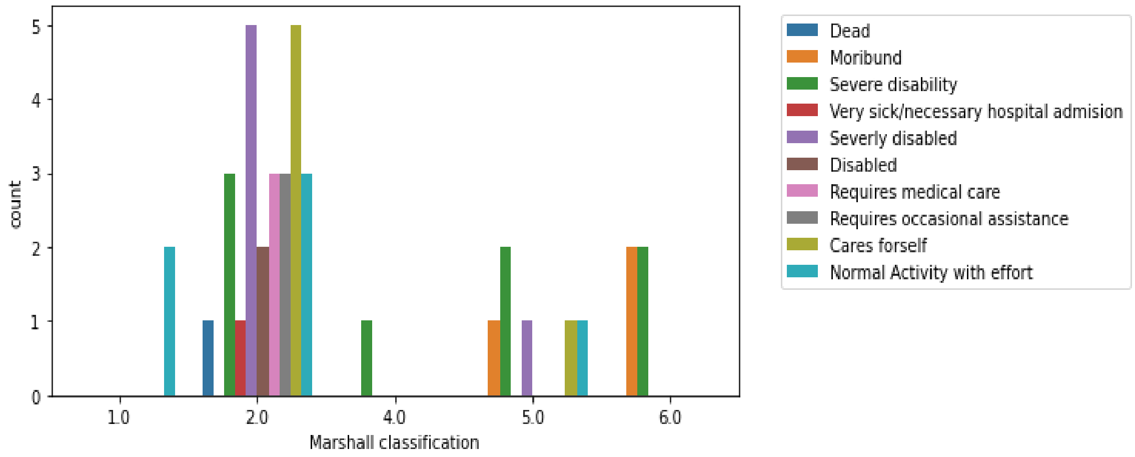


Figure 3.21: Marshall score with KPS

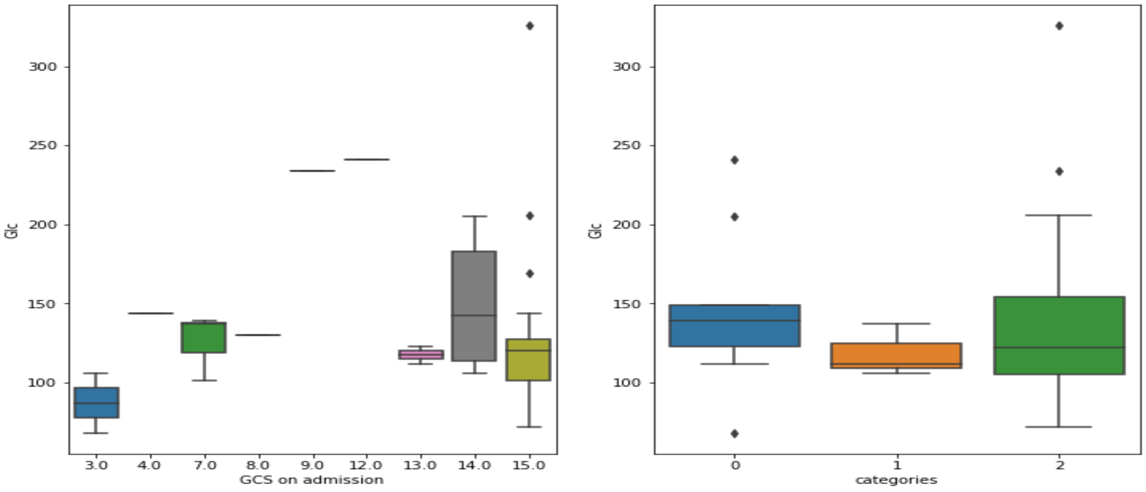


Figure 3.22: Glc grouped by GCS

To sum up, it is obvious that through this visualization of features and exploratory data analysis, we came up with some interesting patterns in data. For example, smaller numbers of interleukins show a progression on the recovery of patients, while marshall and rotterdam scores seem to also be higher for more severe damages. The next step is to continue through models and find out how they estimate the importance of the features.

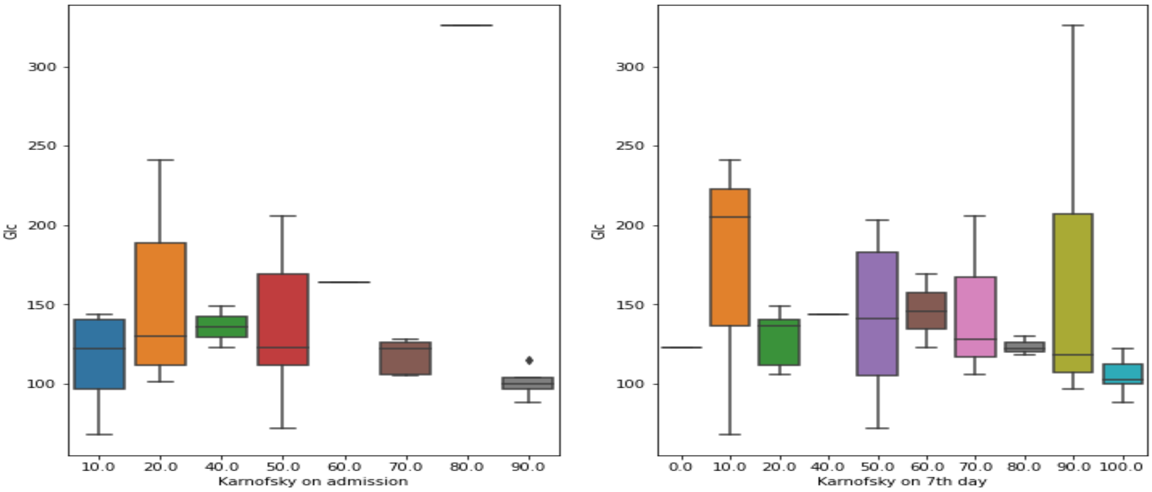


Figure 3.23: Glc grouped by KPS

Chapter 4

Experiments

In this section, we present the methods and evaluation metrics used grouped by the different targets. All the models were trained and evaluated using Leave One Out cross validation [17]which was repeated 3 time for each one and calculated the mean of each score. Also, we included stratified dummy classifier for each target for comparing its performance against other algorithms and so check if selecting categories by random can have same performance. The results of the dummy classifier were low in terms of accuracy and f1_micro(approximately 0.4), which indicates that we can successfully predict the outcome of traumatic brain injury better than randomness. We structured this section by presenting the experiments for each of the 3 indicators in different subsections. The main requests that we need to clarify if we are able to predict with a high accuracy the outcome of patients with TBI and identify the features that have the most predictive capacity.

Glasgow Comma Scale

Glasgow Comma Scale is the most common indicator of patient's status used in other studies. However, defining categories based on GCS produces a heavily imbalanced dataset in our case, since there exist 25 out of the 39 patients with GCS 15. We could try and balance the dataset, but since it is used for healthcare purposes, we need to be accurate and thus we cannot use any up sampling techniques. Therefore, it makes sense for the models to learn and predict the dominant class as it is the safer and most frequently appearing option. Hence, instead of evaluating the model on accuracy, we decided to use metrics like f1_micro, sensitivity, specificity. This metrics are used by machine learning engineers, when they have to deal with imbalanced datasets since they do not take into consideration only the correct predictions.

Table 4.1: Models Evaluation for GCS

Models	F1_micro	Sensitivity			Specificity		
Logistic Regression	87%	100%	100%	25%	50%	0%	100%
kNN	83%	83%	100%	0%	0%	0%	75%
Random Forest	77%	100%	100%	0%	0%	0%	100%
Decision Tree	54%	66%	66%	75%	0%	0%	75%
Extratree	61%	66%	66%	25%	0%	0%	25%
Catboost	67%	100%	100%	0%	0%	0%	100%

We will talk more specifically about these metrics later. So, the table bellow presents the performed algorithms and their evaluation scores:

One of our primary goals is to find the predictive power of the given features. Hence, feature importance was computed in all of the performed algorithms using coefficients(for logistic regression) or feature importance attributes provided by the scikit-learn Python library[18]. The models differ on which features they mostly rely on for their prediction probably due to variations in computation techniques, however we can observe some common patterns. The features listed bellow were included in the top 10 most important features in almost every algorithm: IL-10, IL-6, tSHAH, Marshall Classification, Age, GCS on admission, KPS on admission and PLTs. It is very pleasant that most of the features of interest are contained in the top 10 list. Also, our experiment confirms age as an important predictor in the recovery procedure, conclusion which converges with the results of many other studies.

Moreover, since we have to deal with an imbalanced dataset, our evaluation metric will be f1_micro score. This metric takes into consideration both the number of prediction errors and the type of errors that our model makes and is calculated as the harmonic mean of precision and recall. Since our dataset mostly contains patients with mild traumatic brain injury, the safest option for prediction will be to predict mild. That is what our model does. So, imagine that our test sample contains mostly patients at mild class and the model predicts mild, the accuracy will be very high, but that is not the optimal since we want to check if it classifies correctly every class. F1_score is the optimal evaluation metric as it checks for the proportion of falsely classified classes. Also, sensitivity or recall shows out of the actually positive values how many the model succeed to find. High sensitivity indicates that the model performs well in classifying positive cases, while high specificity indicates higher value of true negative and

Table 4.2: Models Evaluation for KPS

Models	f1_micro	Sensitivity	Specificity	Accuracy
Logistic Regression	71%	66%	100%	75%
kNN	58%	33%	40%	61%
Random Forest	74%	0%	80%	75%
Decision Treet	76%	66%	60%	54%
Extratree	71%	0%	66%	74%
Catboost	77%	33%	100%	65%

lower false positive rate. Note here that recall is also taken into consideration when calculating f1_score.

By looking on table 4.1, we can see that the best model based on f1_score is the Logistic Regression, followed by kNN while both achieve a score over 80%. However, higher mean sensitivity and specificity are achieved by Logistic Regression, which confirms with the findings of other studies. We can see that for classes 0(severe) and 1(mild) specificity is 100%. This means that the model can detect all of these classes correctly, even if this results in some false positive values.

Karnofsky Performance Scale

As mentioned before, Karnofsky indicator produces 10 classes which describe the status of the patient. Further information about what each class represents can be shown in 3.6 of chapter 4. However, due to the small size of the dataset and the fact that the classes as sparse, the model was poorly performing for this task. So, we decided to use binary classification and split the data based on whether the patient can take care of himself or not. This split results into a balanced dataset where 20 patients are able for selfcare and 19 are not. The threshold here is the Karnofsky Performance Status value of 50. Thus, we can proceed in the algorithms summary again shown in Table 2.

Similarly to the procedure followed for Glasgow Comma Scale, we calculate feature importance for Karnofsky Performance Status. The results indicate that IL-10 both on the first and seventh day were important in many algorithms, along side with age, tSAH, Rotterdam score, Hemoglobin, ASDH and IL-6. Again, the predictors of interest are strong predictors even though we perform experiments with a different target variable.

Table 4.3: Models Evaluation for GOS

Models	f1_micro
Logistic Regression	49%
kNN	48%
Random Forest	45%
Decision Treet	37%
Extratree	38%
Catboost	48%

Table 4.2 sums up the results of the algorithms run with KPS categories as target. In these experiments we have a balanced dataset, so we can use accuracy as an evaluation metric. We can see that the best f1_micro was obtained by Catboost. However, this algorithms seems to have poor performance on other metrics. The most balanced algorithm in all metrics seems to be once again Logistic Regression.

Glasgow Outcome Scale

Finally, our last target variable is Glasgow Outcome Scale. Using this indicator, lets as define 5 categories based which show the scale of the damage on patients.

Once more, for this experiment most important features are IL-10, IL-6, tSAH, Glc, age, PT and Marshall classification. Although, the results might not be as good as with other targets, an interesting observation is that IL-10 on seventh day shows up on the top 2 most useful predictors. As we can observe the results of this experiment, we can see that we get poor performances compared to other targets. This is probably occurring due to the fact that we do have 5 categories and we do not contain many registrations on each of them. Thus, the model cannot properly learn patterns of data and it is not able to classify them correctly. Similarly to the GCS experiment, there is no point in using accuracy in this experiment so we rely on f1_score. Therefore, the best model turns out to be Logistic Regression.

Select K-Best algorithm:

Since one of our goals is to find out the features that contribute most to the output, we can use the select k-best algorithm which extracts best features based on a scoring function. Our problem is a classification one, so we use chi function which uses the chi square test and shows which predictors had the highest predictive power on the outcome. Also, we performed experiments with mutual info gain as a scoring function. This function calculates the statistical dependence of 2 variables. For all the possible targets IL-6 on the first day is leading the scores, followed by IL-10 on the first day and Karnofsky on admission. Some other common predictors with high scores are age, tSAH, Marshall Classification, PLTs and Midline shift.

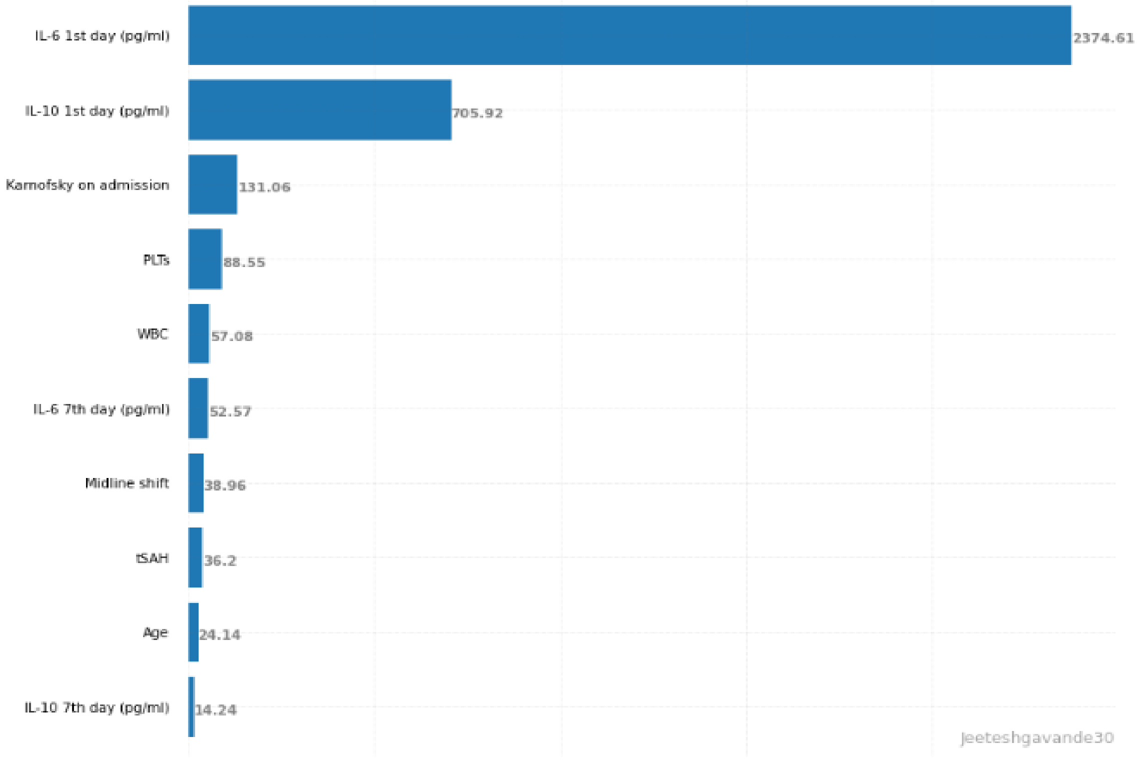


Figure 4.1: Select 10 Best features

Chapter 5

Conclusion

At this point of the study, we have made a clear point on the importance of the study, analyzed each of the collected features, described the process of transforming the data to be compatible for the models and presented the algorithms used to support our hypothesis and make predictions. Hence, we can now move forward into commenting on the results and future plans to improve the study.

5.1 Summary of work

In short, this study presents 6 different models associated with 3 different indicators of the patient's status, while pointing out the importance capacity of features for each algorithm. It is a serious manner to be able to know on which features the machine learning models rely mostly on while making predictions, as it can help clinicians to give more focus on them when considering the medical interventions on the patients. Moreover, we should consider that machine learning models can identify patterns and correlations which humans might note be able to identify. Therefore, the development and evaluation of prognostic models can not only help predictions but motivate or confirm research studies on the association of variables with the output. This work makes use of a unique dataset which contains some commonly used predictors along side with innovative ones and evaluates the performance of the models. Thus, our goal is to present the predictive power of innovative predictors for which there exist some references but are not yet widely used in major models.

In our study, we make use of the 3 indicators (Glasgow Comma Scale, Karnofsky Performance Scale, Glasgow Outcome Scale) each one of which is associated with a different dataset where patients are distributed in multiple categories based on their capabilities of performing daily activities. The variance in the datasets is due to the fact that each predictor has a unique categorization system of the patients status. As mentioned earlier, each of these distributions resulted into unequally balanced datasets. The fact that we had our hands on a small sample along side with the imbalance in the categories made it difficult to get very high accuracy, since the models did not have much data to learn on every category. Despite our efforts to balance the datasets by grouping categories together, the dominant classes were more likely to be predicted than the more rare ones. There are some ways to fix this problem with up sampling data techniques, but since our study is for healthcare purposes, inserting zero-valued samples between original samples to increase the sampling rate was not an option. So, despite the imbalance problem we decided to consider other evaluation metrics such as f1 _score, which do not take into consideration if the prediction was correct on the test data, but rely on the proportion of falsely predicted values. Thus, we perform experiments with 3 datasets and 6 algorithms in order to clarify firstly if the task of prediction the patient's status after TBI is viable and if so, which features/characteristics have the most significant impact on the output.

A common denominator in all of the experiments is that despite the variation in the classification system, Logistic Regression turned out to work pretty well in all cases. It achieved good results on accuracy, sensitivity, specificity and high f1_scores(approximately a mean of 80%), which indicate that we are able to successfully develop prognostic models on patients with traumatic brain injury better than randomness. Furthermore, we can observe some common patterns in the predictive power of features. Although, each experiment consisted of different datasets produced by each status indicator, the interleukins IL-6 and IL-10 were presented as the most important features in most cases either on admission, the seventh day or both. Also, Marshall classification and Rotterdam score were present in many cases, while glucose was only present in cases were we targeted GOS. Scores on admission, like GCS and KPS were listed as helpful but not very frequently. Despite the features of interest, other predictors such as age, tSAH, visible vasal cisterns and Midline shift were also listed as important. Note here, that tSAH appeared in most cases as important and also it is interesting that it is a score that helps calculate Rotterdam score and Marshall classification. These conclusions confirm findings of other studies, since age and tSAH have been mentioned from many studies for their predictive power.

The final conclusion of the study is that there seems to be a strong correlation between the values of interleukins and Traumatic Brain Injury effects which need to be further investigated as it can help to improve the accuracy of existing prognostic models. We can verify that while patients on admission presented high values of interleukins, during their recovery procedure, these values seem to fall over the 7 day period. More specifically, IL-6 and IL-10 on admission appear to have the strongest predictive power in all cases. Remember that early accurate predictions result in a lot of benefits both in healthcare and economic concerns. It is well known that low-middle income countries often suffer problems of filling hospital facilities with patients. Imagine that an early prediction would avoid the overpopulation in hospitals, give doctors time to consider more cases, improve life quality of patients and free medical resources so that they can be used whenever they are necessary. However, since we are trying to develop a model for disease prediction outcome we need to be very accurate. Thus, even though we got some first promising results, we need further internal and external evaluation on larger datasets in order to be sure for the conclusion. We hope that our study can motivate other researches to look into the usefulness of these factors and who knows maybe we can update the predictive power of the already existing models or define new ones.

5.2 Future work

The fact that the obtained results are promising and prove our hypothesis that innovative features can have strong predictive power, motivates us to continue expanding our study on a larger scale. At this moment, the bioethics committee has approved our study and we are currently waiting to get access on much more data on patients with Traumatic Brain Injury. Since we get our hand on these new information, we aim to rerun the experiments, identify the most accurate model and have more data to generalize and further support our findings. Furthermore, there are some claims from related works which suggest that the use of multi modal learning can improve accuracy. Multi modal learning uses the joint representations of different modalities. In our case we can include the given dataset, analyze images from CT scans through computer to find any anomalies and take into consideration a doctor’s early description of the patient’s status. Afterwards, we aim to develop a website that is more user-friendly, which not only clinicians but also researchers, would be able to consult and perform experiments and maybe come up with some new findings that would alter the idea we had about the influencing factors of Traumatic Brain Injury. Finally, more accurate predictions offer a chance to optimize management of medical resources and on time interventions to save lives or improve life quality of patients. However, there are few limitations when experimenting with medical data like the difficulty in obtaining registrations along with the fact that prediction of disease outcome requires the development of very accurate prognostic models, trained in large samples and with high external and internal validity. Also, medical confidentiality can often slow down the development of researches. That is why a web-site that keeps up with bioethics by hiding the personal information of patients and offers the chance to experiment with different predictor values is a need.

Bibliography

[1] George L Sternbach. The glasgow coma scale. *The Journal of emergency medicine*, 19(1):67–71, 2000.

[2] Tom McMillan, Lindsay Wilson, Jennie Ponsford, Harvey Levin, Graham Teasdale, and Michael Bond. The glasgow outcome scale—40 years of application and refinement. *Nature Reviews Neurology*, 12(8):477–485, 2016.

[3] Hamid Reza Talari, Esmaeil Fakharian, Nooshin Mousavi, Masoumeh Abedzadeh-Kalahroudi, Hossein Akbari, and Sommayeh Zoghi. The rotterdam scoring system can be used as an independent factor for predicting traumatic brain injury outcomes. *World Neurosurgery*, 87:195–199, 2016.

[4] Allen W Brown, Christopher R Pretz, Kathleen R Bell, Flora M Hammond, David B Arciniegas, Yelena G Bodien, Kristen Dams-O'Connor, Joseph T Giacino, Tessa Hart, Douglas Johnson-Greene, et al. Predictive utility of an adapted marshall head ct classification scheme after traumatic brain injury. *Brain injury*, 33(5):610–617, 2019.

[5] Amanda Ross and Victor L Willson. One-way anova. In *Basic and advanced statistical tests*, pages 21–24. Springer, 2017.

[6] Roshani K Prematunga. Correlational analysis. *Australian Critical Care*, 25(3):195–199, 2012.

[7] Anthony Marmarou, Juan Lu, Isabella Butcher, Gillian S McHugh, Nino A Mushkudiani, Gordon D Murray, Ewout W Steyerberg, and Andrew IR Maas. Impact database of traumatic brain injury: design and description. *Journal of neurotrauma*, 24(2):239–250, 2007.

[8] Bob Roozenbeek, Hester F Lingsma, Fiona E Lecky, Juan Lu, James Weir, Isabella Butcher, Gillian S McHugh, Gordon D Murray, Pablo Perel, Andrew IR Maas, et al. Prediction of outcome after moderate and severe traumatic brain injury: external validation of the impact and crash prognostic models. *Critical care medicine*, 40(5):1609, 2012.

[9] Sharhokh Yousefzadeh-Chabok, Anoush Dehnadi Moghaddam, Ehsan Kazemnejad-Leili, Zahra Saneei, Marieh Hosseinpour, Leila Kouchakinejad-Eramsadati, Alireza Razzaghi, and Zahra Mohtasham-Amiri. The relationship between serum levels of interleukins 6, 8, 10 and clinical outcome in patients with severe traumatic brain injury. *Archives of trauma research*, 4(1), 2015.

[10] Robson Luis Amorim, Louise Makarem Oliveira, Luis Marcelo Malbouisson, Marcia Mitie Nagumo, Marcela Simoes, Leandro Miranda, Edson Bor-Seng-Shu, Andre Beer-Furlan, Almir Ferreira De Andrade, Andres M Rubiano, et al. Prediction of early tbi mortality using a machine learning approach in a lmic population. *Frontiers in neurology*, 10:1366, 2020.

[11] Syed M Adil, Cyrus Elahi, Dev N Patel, Andreas Seas, Pranav I Warman, Anthony T Fuller, Michael M Haglund, and Timothy W Dunn. Deep learning to predict traumatic brain injury outcomes in the low-resource setting. *World Neurosurgery*, 2022.

[12] Ahmad Abujaber, Adam Fadlalla, Diala Gammoh, Husham Abdelrahman, Monira Mollazehi, and Ayman El-Menyar. Prediction of in-hospital mortality in patients with post traumatic brain injury using national trauma registry and machine learning approach. *Scandinavian journal of trauma, resuscitation and emergency medicine*, 28(1):1–10, 2020.

[13] MRC Crash Trial Collaborators et al. Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. *Bmj*, 336(7641):425–429, 2008.

[14] Cheng-Shyuan Rau, Pao-Jen Kuo, Peng-Chen Chien, Chun-Ying Huang, Hsiao-Yun Hsieh, and Ching-Hua Hsieh. Mortality prediction in patients with isolated moderate and severe traumatic brain injury using machine learning models. *PloS one*, 13(11):e0207192, 2018.

[15] Ewout W Steyerberg, Nino Mushkudiani, Pablo Perel, Isabella Butcher, Juan Lu, Gillian S McHugh, Gordon D Murray, Anthony Marmarou, Ian Roberts, J Dik F Habbema, et al. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. *PLoS medicine*, 5(8):e165, 2008.

[16] Andrew T Hale, David P Stonko, Jaims Lim, Oscar D Guillamondegui, Chevis N Shannon, and Mayur B Patel. Using an artificial neural network to predict traumatic brain injury. *Journal of Neurosurgery: Pediatrics*, 23(2):219–226, 2018.

[17] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27(5):1413–1432, 2017.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.