Data analysis and prediction with medical data and the resulting opportunities for companies in the medical industry

On the basis of an example analysis

Submitted to Dr. Daniel Dan

Paul Zeileis

11831822

Vienna, 31.05.2023

Table of Contents

List of Figures

Abstract

This thesis aims at defining the opportunities and advantages a company in the medical sector can acquire with the aid of data science. Among the resulting possibilities, the focus lies also on the different methods, approaches and subfields of data science as well as positive examples and ethical considerations.

Data collection, processing, and evaluation are just logical steps as more and more information becomes accessible. Every year, more data is produced than in the previous year and the methods with which this data can be examined are also growing rapidly.

Besides all these theoretical aspects, a data collection and subsequent analysis of data from a company in the medical field will be carried out for a better understanding of the topic covered. The experiment is based on a machine learning algorithm.

This self-conducted data research and the following analysis are a step forward to understand what in general data science is about and how it can be used to give advantages to companies, research institutions, non-profit organizations and more as well as to the patients themselves. Also, possible third parties like insurance companies for example can profit from such findings. The goal of every field of science is to bring the public and individual groups further forward.

The usage of data science has already shown remarkable results in the medical industry as well as in all other fields where it is applied.

Especially the field of medicine looks promising to be one of the biggest beneficiaries of these modern methods.

1 Introduction

According to statistics, in order to optimize every little uncertainty in all internal processes, 53% of companies make use of data science (Columbus, 2017). In general, the gathering, evaluation, and interpretation of big data sets, seems to get increasingly important when it comes to customer related information (Blei & Smyth, 2017).

One can imagine that such processes, like investigating huge data sets and draw useful conclusions out of it is a costly approach to gain maximum insight, especially when companies have thousands or millions of customers.

Companies therefore use science to maximize their profits and gain advantages over competitors. But, when it comes to the medical industry and medical patients´ data, through proper data evaluation, both sides can benefit from it (Banerjee et al., 2017). Through this, personalized medicine and specific individual treatments can be achieved, not only when it comes to acute medicine or diseases, but also in terms of health care (Terkola et al., 2017).

Data evaluation particularly has a special meaning in the medical industry because resulting opportunities for companies are strongly connected to the benefits the patients gain from the process. This makes the topic even more important and interesting.

In this thesis it is shown how smaller companies which do not have huge monetary support are also able to conduct useful analyses. Data is everywhere, exceptionally in medical facilities on the basis of all the processes and treatments that include several medical values. Based on an on-hand case study where medical data of a medical institution is evaluated, it is demonstrated what opportunities a company, that usually does not investigate patient´s data, can gain. Of course, such processes also consist of challenges as well like the anonymization and security of the data (Armstrong, 2017). Whenever there is a challenge there is always a resulting opportunity and in this case it is a better coordination and handling of important medical findings from which the patient as well as the company benefit.

Besides the case study to literally show what such process is about and to give a better understanding, several questions are dealt with in this thesis.

First, there are a few questions that must be answered to give a basic understanding. What are the different methods and subject areas of data science? What is data

prediction? How can a company find the perfect way to start off with their data collection? The next step is implementing the found aspects to the strategy. In which ways can medical companies benefit through data science? How far goes data research of big companies and do they have any boundaries or ethical considerations? With use of these questions and the conducted on-hands case study the goal is to prove that analysis and evaluation of data pay off every time it is carried out, along with raising awareness of modern procedures like data science as well as data prediction, in particular the Random Forest model.

The first part of the thesis after the introduction, the literature review, lists various journal articles, research chapters, and academically relevant literature related to data science. The core statements as well as important definitions and classifications are described so that a better understanding of the material is possible. All the individual subchapters of this part have a significant meaning for the thesis and are arranged in a more and more in-depth way.

The next chapter, methodology serves as an overview of the strategies and general approaches used in this thesis. The methods for the literature review and the experiment are explained. The method of the experiment is described, and limitations and difficulties are discussed.

In the fourth chapter, data analysis and results, and also all aspects related to the analysis are covered. Started from an overview of the data, through the analysis itself to the results and interpretation. At the end of the chapter the benefits for the medical company as well as the recommendations that can be given are reported. Another point that is considered is the further research that would be possible on the basis of this thesis.

In the last chapter, summarizing conclusions are drawn and a look into the future is also given. Following is a list of abbreviations which also serves as an explanation as well as the bibliography and the appendix.

2 Literature Review

2.1 Data Science

Data science is a twenty-first-century buzzword and umbrella term for many different types of work (Juavinett, 2020). However, mostly it refers to work processes that deal with organizing, analyzing, and structuring data. The collection and storage of data also play a very important role in this scientific field. Juavinett´s work is of importance for this thesis as she also deals with the connection between data science and medicine.

Data science has attracted a lot of attention, promising to turn vast amounts of data into useful predictions and insights (Blei & Smyth, 2017). Their work focuses mainly on the different components of data science, which of course is of crucial importance for this thesis. It combines programming skills with knowledge of mathematics and statistics. Of course, industry-specific expertise is also required, depending on the area of application. The benefit of this is that it allows useful knowledge to be generated from large quantities of data that are usually far too large for a human to sift through. For this reason, the term Big Data is often being used. But of course, the same is possible with smaller and more manageable datasets. Big data offers unique analytical opportunities to reveal patterns that may otherwise have gone unnoticed (Ricker, 2018). On the grounds of this, this field of science is interesting to so many. According to Sieber & Tenney (2018) the hype surrounding data science affects all areas of social life and shapes the way researchers deal with digital data.

A new era in the handling and processing of data. The aim is to describe, predict, and ultimately optimize or learn. Before the different methods can be applied, however, the data must be collected.

As Baker & Henderson (2017) wrote, there are huge amounts of data available, to stay in the technical jargon, Big Data. Most of this data is stored in data collections, so-called databases, which do not offer public access (Landgraf, 2018). That is why the main operators are mostly large companies. Large companies not only have the monetary resources, but they also have comprehensive data on customers. Providers of apps or digital goods are particularly well equipped with data. This is because most apps require access to sensitive data on users' devices. But data can be obtained in many ways. In today's digital world, everyone leaves their mark in the form of data,

intentional as well as unintentional. Whether grocery stores or medical facilities, data is collected everywhere that can be evaluated. Among digital and analog data gathering there are several methods to gather information without investigating these huge databases that were already mentioned. Especially popular are surveys, interviews, observations, online tracking, and social media monitoring according to Harvard Business School Online (2021). In this way, data can be generated that is tailored to the purpose of the data collection and thus has an advantage over data collected from large databases.

Parsons (2017) explains the definition of data evaluation, namely the systematic and empirical analysis and the consequent conclusions that can be drawn from data. His work is especially important for this thesis as he describes, along with the definition, the different methods of evaluation and declares how to correctly evaluate to maximize effectiveness and quality. Another crucial step in the evaluation is to separate the gathered information into two sections, data that can be used and data that cannot be used in the process. Venezia (2018) explains how to filter out the relevant data using tools that can facilitate the process. For this step it is important to choose the proper evaluation method for the specific research approach.

Shu (2020) stated that large and complex data sets displayed in tables often disguise certain patterns and systematics. Her work is of great significance for this thesis in terms of the conducted analysis of a medical institution and the resulting evaluation and visualization.

2.2 Data Prediction / Predictive Analytics

Cetintemel (2018) describes that predictive analytics refers to the application of different analytical procedures. According to the author, this is about data-driven modeling, mining, and learning from historical data to make predictions about missing or incomplete data values or patterns.

Data prediction is often used in connection with predictive analytics, and it deals with the combination of prediction tools and already collected data. The basic idea is to make predictions as accurately as possible. Like other methods and procedures, data prediction is an integral part of data science. Terms, such as data prediction and predictive analytics, are the building block for this thesis as this is the basis of the experiment. It is important to mention, that there is a difference between data

prediction and predictive analytics. Data prediction is when machine learning attempts and statistics are used to process historical data to predict events in the future. Predictive analytics, on the other hand, is when patterns and trends can be filtered out from data to help with future decisions. Fairly accurate data forecasting can therefore bring benefits in many industries and sectors, such as finance and healthcare. To bring predictive analytics in connection with medicine the work of Michard & Teboul (2019) indicates similar paths we took in this thesis. In their journal article they describe the opportunities that emerge through the enormous amounts of clinical data fabricated by electronic medical records and physiologic monitors. The two authors clarify that high performance computers are capable to detect clinical deteriorations through applying machine learning and predictive algorithms. The authors also address the random forest model which is explained in another chapter. Furthermore, the publishers write that scientists have succeeded in creating a predictive algorithm to forecast the risk of mortality of patients. Having the certainty that such procedures work, the findings give support for this thesis. This algorithm is called ICU learning algorithm (ICULA) and further explanations as well as more details at the procedure are declared by Pirracchio et al. (2015). The ICULA is a so-called ensemble machine learning method that makes use of many different learning algorithms to achieve excellent prediction results. The algorithm works with so-called severity scores and generates much more accurate forecasts than any other similar method using severity scores (Pirracchio et al., 2015). Breakthroughs like this inspire even more widespread use and development of such algorithms.

2.3 Machine Learning Algorithms

Machine learning algorithms are computer procedures that use pattern recognition to produce the desired result (Liu et al., 2019). Their work mainly revolves around the usefulness and application of machine learning algorithms which makes it highly essential to this thesis. Liu et al. (2019) further state that to take full advantage of the information accessible analysts must be able to examine and visualize large amounts of data in a timely manner. The main objective is to generate artificial knowledge based on empirical values. As its name implies, the machine learns by example and can later apply what it has learned in many ways.

Mullainathan & Spiess (2017) declare that that the final breakthrough of machine intelligence has both statistical and computational reasons. In their article, they discuss how machine learning algorithms became possible in the first place and how the shift from procedural learning to empirical learning played a role in their creation. For this thesis, the work of Mullainathan & Spiess (2017) has a tangible significance since the analysis and prediction in the experiment are performed using a machine learning algorithm.

“Machine learning algorithms are motivated by the mission of extracting and processing information from massive data which challenges scientists and engineers in various fields such as biological computation, computer vision, data mining, image processing, speech recognition, and statistical analysis“(Zhou, 2015). Furthermore, he explains that clustering, classification, feature selection, ranking, and backstepping are all tasks of machine learning algorithms. There are also three different types of algorithms, namely unsupervised machine learning algorithms, supervised machine learning algorithms, and a mixture of the two, semi-supervised machine learning algorithms. He discusses the types of machine learning algorithms in his paper. For the analysis afterwards it is advantageous to understand the diverse types.

According to Saraswat (2022) the goal of supervised learning is to set up an exact model of the distribution of class identifiers in terms of predictive characteristics.

The core of supervised machine learning algorithms is therefore to command the machine or computer to use this empirical data to solve a specific problem. Or, as in the experiment later, to gain more knowledge that can then in turn be used to optimize processes and generally create more clarity about the data. He explains subsequently that these algorithms usually perform classification tasks in the world, such as distinguishing regular emails from spam.

This research heavily relies on the work of Balakrishna and Anandan (2020) on the application of supervised machine learning algorithms in disease prediction. Their work highlights the importance of such algorithms in predicting diseases and how they relate to the decline in mortality rates for some diseases. A good knowledge of the potentials of these algorithms becomes possible by including their work. The best-known machine learning algorithms include linear regression, logistic regression, support vector machines, random forest, naive bayes algorithms and so on.

2.4 Random Forest

Random forest is a supervised machine learning algorithm that mostly deals with regression and classification tasks. According to Vens (2013), random forest is an ensemble of random decision trees that makes forecasts by performing a combination of the forecasts of the separate trees. She also expresses that this method can be applied to perform predictions on nominal as well as numeric target attributes. In the case of nominal character traits, this is called classification. When it comes to numerical character traits, it is called regression. Furthermore, she addresses the characteristics of the random forest model and how different random forests vary in the randomness of the tree formation process.

How exactly the supervised machine learning algorithm works is explained by Sammut and Webb (2017) in their research chapter. As already mentioned, this learning method relies on decision trees as a base classifier. The two authors explain that one of the most important aspects is that decision trees are not pruned after creation, allowing for overfitting to one's own sample of data. At each individual branch in the decision tree, the decision of which feature to split is restricted to a random subset of size n from the total set of features, further diversifying the classifications, according to the authors. Finally, this random subset is re-selected for each decision.

The random forest technique therefore has a lot of advantages over other algorithms that deal with classification and regression. One of the main advantages is that it is able to generate and train a large number of decision tress at the same time while being highly efficient. Because of this high accuracy and rapidity, the random forest algorithms are particularly suitable for complex analysis tasks. Each tree deals with an individual subset of the data. From many individual decision trees, a so-called forest emerges. In this way it is possible to make much more explicit predictions than predictions from single decision trees. Another big benefit is that the algorithm is able to cope with missing data. Due to the many trees in the forest, it does not matter too much if single trees do not give perfectly accurate results. On the basis of the stated potentials, it is clear that the random forest algorithm is a beneficial tool for many data scientists and finds application in more and more different fields and branches.

2.5 Successful Examples from the Medical Sector

According to several academic research papers and journal articles, like the Irish journal of medical science and health data science, there seems to be an increased application of medical data evaluation in Asia compared to Europe. Zhang (2018) discusses the medical research and the different approaches in China. Her findings verify that by executing data evaluation, an organization is able to elaborate a competitive advantage and the patients' needs are handled more efficiently. Further she was able to prove that a country can benefit through constant monitoring in terms of making trends of diseases visible as well as ease the process of medical research (Zhang, 2018). In many fields, especially in medical research, the use of visualization in order to evaluate trends and patterns is often applied. With the help of the visualization process it is a lot easier to understand complex issues, which enable all kind of researchers to notice patterns and draw conclusions that would not be immediately observable from raw data. Simplifying medical research is as important as visualizing it. This thesis, which tries both, simplifying medical data and visualize it afterwards is a perfect example of this strategy. In this way it is possible and uncomplicated to share and provide findings to a larger audience. This can result in people, companies and organizations making better and thoughtful decisions in terms of precautionary actions but any type of choices in general. Cost savings are a benefit too. Particularly in the medical field these approaches are extremely important whether the concept is applied on a large or a small scale.

Tsuji (2020) explains how the Japanese government deals with big data in terms of handling the aging society in Japan. The author´s principal concern in his work is the supply of medical services for the older section of the country's population. In detail it is about the difficulties these senior citizens must deal with regarding the financial challenges that come along with getting older. Because the average age is rising in Japan it is more important than ever to provide healthcare possibilities that is accessible for as many people as possible.

To improve the accessibility to preventative examinations, such as routine check-ups and screenings is one important strategy. This way potential health issues can be identified in an early stage and may help avoiding late occurring issues. The Japanese government additionally calls up on the population to maintain a healthy lifestyle. Examples that are mentioned are a balanced nutrition as well as doing sports on a

regular basis. The matters stated can also help to avert illness in addition to the given data based methods. Overall, his work once more demonstrates how data analysis and prediction can have significant impact on a broad scale.

By applying the most recent tools in data science and machine learning it is easily possible to help individual people, groups, or whole industries as it is shown. The knowledge is just latent in huge datasets and needs to be filtered out by using exactly these techniques. Tsuji (2020) addresses data prediction on the highest level since an incredibly large amount on data is on the Japanese governments hand.

Borgheresi et al. (2022) report about an Italian project called Navigator, where precision medicine in the field of oncology is made more predictive and therefore also more preventive and personalized. The data used for this project contains among the conventional medical data especially the results from all kinds of imaging methods (CT & MRI). With the help of open source databases that contain the mentioned medical information a digital patient model, which provides reliable forecasts concerning cancer diseases could be obtained, so the authors mentioned above.

But not only on the scale of entire countries, data evaluation can make a huge difference and give crucial insights for an organization and all kind of stakeholders including patients as well, as Banerjee et al. (2017) stated. However, this thesis shows in general how the evaluation and analysis of data can be crucial to businesses across all sectors. An organization or company can understand many individual areas like consumer behavior, market trends or the own processes and operations way better by using these modern sciences. Or as in this case a medical institution, their own patients and the hidden relations between certain values and conditions. Companies are enabled to customize their services and products to better address the needs and wishes of customers and clients collectively. Additionally, data scientists are able to assist businesses to optimize internal procedures like cutting waste or use resources more efficiently. The materials, employees and money are meant by resources. Of course, the economical handling of resources is desired also by medical businesses, only that here special aspects come in addition, namely the well-being of the patient. Through all the examples above it can be seen how well this can work out when the right techniques and procedures are applied.

2.6 Criticism of Data Science

This subchapter deals with the criticism surrounding data science and will examine the negative aspects of it and big data to illustrate other facets as well. Authors Carter & Sholler (2016) describe in their work exactly what they consider to be the biggest problems in the face of data science.

They state that both data science and big data have gained a lot of attention in recent years, but it has also been accompanied by a lot of criticism. This refers to questions concerning the efficiency and effectiveness of data science. Also, implementations and consequences are meant here.

It is put forward that supporters of data science and big data have not taken privacy risks into account. Also, the further authorization of governments presents a risk.

Theory and more conventional approaches are often ignored because the subject is about if a statistical model can produce reliable outputs or not. The reason for these results is mostly secondary.

Critics claim that traditional approaches are sometimes more effective than data science and big data approaches. These approaches continue to depend on the opinion of experts in the respective field to review findings and show what the data does not explain.

The two writers also state that, for example, supporters of these approaches are likely to neglect the expertise of the researcher rather than justify the data itself. Many people have the opinion that the output computer work consists of facts and is not interpretative. The reason is that a model provides results, alleged facts, and only when the researcher tries to interpret these facts does the interpretative process begin. Big data and data science are therefore not determined facts but consist of information collected in the course of the researcher's objectives and plans.

According to the two authors the strategies that information researchers apply to do their work can include another level of transmission. The characteristics of advancements can serve as guidelines or benchmarks for specific operations and influence the way information is collected, processed, and analyzed.

Analysts have found that social variables, such as peer connections, impact employee’s decisions when working and association with innovation utilize and gathering of people’s desires (Carter & Sholler, 2016).

Further points of criticism relate to the procurement of data. The authors believe that this restriction on the procurement of data will build a novel form of digital separation. The so-called “Big Data rich” and the “Big Data poor” as stated by the authors. Because of that, researchers who work for large social media corporations are better positioned, because full and comprehensive data sets are only available for researchers who work in companies or the ones who have the monetary resources to purchase data from them (Carter & Sholler, 2016).

It is important for a scientific paper not only to deal with the good sides and advantages of a topic but also to discuss openly the negative sides and disadvantages.

2.7 Ethical Considerations in Data Science

The many facets of ethical and responsible data handling are explored. This includes the collection, storage, and use of data, which play important roles in the operating with data. The chapter aims to provide an overview of all the ethical issues in data science and what needs to be considered. Floridi & Taddeo (2016) determined that data ethics builds on computer or information ethics, but over time has established itself as its own part. Furthermore, they highlight the importance of data ethics but point out the complexity and challenges it brings with it. The authors believe that data ethics should be processed as so-called macro ethics. Macro ethics is understood as an overall framework that avoids unexercised approaches but addresses the ethical implications and impacts of data science in a complete and comprehensive framework according to the above-mentioned authors. They acknowledge that data ethics can only offer good solutions for society if it is carried out as a macro ethic.

Richterich (2018) addresses in her book chapter very specifically the monitoring of public health in relation to big data. The author's focus is on data mining with regard to social media and all relevant ethical components. It is apparently normal for many people to post about their health on social networks and to go into detail about it. Thus, data ethics is not only a relevant topic for those who collect or store data, but also for the people who virtually put their own data into circulation. Although people are often advised not to do so, lots of people present their own personal and confidential data lightly. Often it is underestimated what is possible with such kind of data.

In all fairness, it is difficult to shed ethical light on a subject when new opportunities arise regularly. With all new methods that data science brings, new ethical approaches and ways of thinking should be considered. In any case, this topic will become increasingly critical in the next few years as data science is developing very rapidly. Therefore, it is important that data ethics evolves at a similarly fast pace.

With the rapid progress concerning artificial intelligence technology data ethics is most likely to become one of the most important ethical subtopics.

3 Methodology

3.1 Research Design

In this thesis a mixed-methods approach is used. Combining qualitative and quantitative techniques for this study has a major advantage over other research designs. Both parts have a key role, and the thesis is split into them to achieve the best results. A comprehensive literature review was carried out to get deeper understanding of the subjects of data science, data prediction, machine learning, and a specific algorithm of machine learning, random forest.

This is the qualitative part of the mixed methods approach and serves as a foundation for understanding the covered topics. The literature review begins with fundamental definitions and concepts, progressively moving on to more complicated applications, and concluding with a detailed explanation of the random forest method. Additionally, the diverse benefits and opportunities that emerge are explored. The potential benefits of using data prediction, machine learning algorithms and the random forest model are addressed. With the experiment and the resulting analysis and interpretation, which make up for the quantitative part, the goal is to illustrate all the theoretical components of the literature review. To get the best results, the machine learning algorithm random forest is utilized to forecast and classify medical values on the basis of the datasets provided by a medical institution. As the next step, conclusions and helpful interpretations are drawn out of the analysis. This includes preparing the results in a way so that the company can make use of them on the subject of understanding connection between various medical values better. In this way it is possible to provide more accurate treatment for individuals.

A detailed investigation of the research areas was made possible by the usage of a mixed-methods technique in this study, leading to a more in-depth comprehension of the subject. While the quantitative part supplies actual data and useful applications, the qualitative aspect gives a theoretical foundation. Every project that involves data analysis must start with the acquisition of data, and for this experiment the provided data is a patient data set was gathered from the year 2018 to 2022. Using existing data compared to collecting new data specifically for this experiment has several benefits, including time and resource savings. It is important to note that the data being used in this investigation has been anonymized. That means that just medical values and demographic data is present in the data set while personal information like name, address and payment information was logically removed. An important part of research using personal data is maintaining the privacy of those individuals. The sample size for this study amounts to approximately 3150. As mentioned before all data was recorded in a five-year period from 2018 to 2022. The dataset primarily consists of demographic information, blood values, liver values, and urine values. Medical information that would not have benefited from this study was also removed, so that all parameters that could be studied are included in the dataset.

This broad data collecting is crucial because it makes it possible to draw insights and conclusions that might otherwise not be obvious. A series of figures and plots are used in this paper to explain the facts and the results as precisely as possible and also to make them visible. Among them are simple diagrams as well as more complex plots like the random forest model. Graphs serve thereby to display the entire population. The literature review thus allows an uncomplicated transition from the basic topics to more specific methods and finally to the application through the random forest algorithm. Overall, the use of mixed techniques in this study enables a more thorough examination of the data, which results in a more informed and insightful discussion of the findings.

3.2 Approach to the Data Analysis and Evaluation

In this subchapter the exact procedure of the experiment is described and explained. In addition, valuable information about the medical company is given.

As mentioned before, the medical data of a group of patients is analyzed in the conducted experiment. The medical facility, which provided the data is located in

Upper Austria but it is undisclosed. It is important to note that this is a facility that offers various testing as well as several treatments and therapies. Most patients come to the cure which has preventive purposes. In contrast to a hospital no acute injuries are treated. The average patient stays between one and two weeks as most of the therapies are designed for this period of. The treatment program is designed for each patient individually. Depending on the patient, different examinations are carried out, although there are a few tests that are applied to everyone. Among them are the small hemogram, electrocardiogram, and urine tests.

The data was collected with the purpose of finding correlations between values in order to gain new knowledge that can, in the best case, be implemented in new strategies.

The experiment was carried out with the help of the R language, and the R Studio IDE, respectively. These terms have a close connection and yet have two different meanings. While R is the programming language, R Studio gives a user interface to it. Thus, all desired operations can be performed and illustrated.

The goal is to get the accuracy with which a certain value can be predicted when the other values are available to the person or company performing the analysis. Thereby important knowledge is gained, from which it can be concluded to conduct or not conduct particular tests or treatments. This way the medical institution might save on examinations that are unnecessary or superfluous in some situations or is able to recognize at an early stage that some specific checks need to be conducted. In addition to the potential for early detection, the company can also save time, employee effort and resources.

3.2.1 Difficulties with the Analysis

Problems and difficulties must be overcome at some point in almost every experiment. In this case, the most difficult thing was certainly to find a suitable method for the evaluation. Since this is a classification task, the initial idea was to use support vector machines for the analysis. The plan was to divide the patients into several classes depending on different variables. The goal should be to show the differences of the found classes and to generate knowledge from them. Over time and after becoming familiar with the available data, the random forest algorithm has

proven to be most promising. The process of finding the optimal method and getting to know the data took a considerable amount of time.

To get to these results a lot of random forests with a lot of variable combinations were created and most of them did not yield practical results. It was challenging to select the values that serve as the basis for several reasons. First, the values selected must stand in some relation to the target value from which the accuracy is predicted otherwise it would not be possible for the medical institution to draw conclusions out of it which can help to optimize certain procedures. In theory there is also the possibility to obtain meaningful results, but the outcome would just not help or generate new knowledge for the company. The second reason is that some combinations of values in the dataset and the selected target value simply did not work to predict the accuracy at an acceptable level.

With help and in consultation with representatives of the medical facility various promising sounding attempts were carried out. The analysis shown in this thesis has the best use for the company as helpful reverse conclusions can be drawn out of it.

3.3 Limitations

It is a law of executing research to state the limitations of the study openly and honestly and to not give high hopes. If research limitations arise, these weaknesses can lead to the study being distorted or misinterpreted, because they have a significant influence on findings and outcomes within the data (Ross & Bibler, 2019). This chapter serves to make the various limitations clear to the reader. Despite the limitations mentioned below, the maximum was extracted from this experiment.

3.3.1 Lack of Data Availability and Diversity

Even if a sample size of about 3150 is not a small group of representatives it is also not big enough to really discover a medical trend in large groups of the population. The results, despite their significance are so to say only relevant for this medical institution because the data provided just presents the average patient of exactly this medical facility. In general, there are several aspects that make this set of data very special and therefore the results cannot be applied to all people. The following points in particular are decisive here, namely gender, age and nationality.

As for the gender it must be mentioned that a different distribution between males and females could lead to other results and therefore other accuracies in the analysis. In this experiment 57% of the data are females while the remaining 43% are males.
The same reasoning applies for the age variable as the average age from the sample size is around 68 years. Other medical facilities that have the same values tested have younger groups of patients. This might lead to different results and accuracies.
Nationality is the hardest part to generalize about as people around the globe live in different climate zones and also the sea level differs a lot globally. Additionally certain diseases are more common in some areas of the world. For this experiment about 85% (2740 out of 3151) of patients are Austrian citizens. The next biggest groups are citizens from Germany and Switzerland, which are not too dissimilar to Austria in terms of conditions. And only 0,005% of the total sample size are people from outside Europe.
This makes it really difficult to generalize the results found in the analysis.

3.3.2 Missing Values

In addition, comes that as mentioned before not all patients are examined with regard to all values. This leads to some empty cells in the dataset which must be handled and for that reason the dataset is even smaller after the data is cleaned.
For some processes this was a problem as it did not produce the desired results. Also, it was tried to fill in the missing values with the help of built-in functions to get around this problem. Unfortunately, due to this limitation, not all the analyses that were planned in the first place could be carried out in the right manner. There is also the possibility to anticipate the missing values using the impute function included in the mice package (van Buuren & Groothuis-Oudshoorn, 2011). But this option would bring along several problems. Thus, the imputation would not only require a lot of time and computing power but would also produce fictitious numerical values as a basis for the analysis.

4 Data Analysis and Results

4.1 Breakdown and Summary of the Sample

This chapter serves as summary of the patient data to get a good overview. The sample is formed of data from 3151 patients collected in a five year time period. There are 38 columns of different values and characteristics in the dataset. Some are incomplete and some are not relevant to this thesis, therefore only about half of the columns are taken into consideration. About the most important ones of them a few key data is given. Before the most important values tested are discussed, it is about the demographic data in order to give a basic idea of what group of patients it concerns.

1793 patients and therefore a total of 56,9% are women while the remaining 1358 of them are men which amounts for 43,1%. The average age is 68,6 years whereby the range is from 13 years to 102 years. 98,7% of patients are from Austria (2746), Germany (300) and Switzerland (72). Although only a small portion is left, patients from all continents are included in the date set, the largest groups of the rest are from Saudi Arabia (6) and the United States (5).

The values that are dealt with the most are bili, GOT, GPT, AP and GGT. For a more detailed explanation see the list of abbreviations.

With help of the summarise function in R studio the mean of certain values can be calculated to do an exploratory analysis. A variety of options are available in the summarise function. For example, the range in terms of maximum and minimum values per category as well as the median can be calculated. Several other options are also possible by its use. To operate with the summarise function the package dplyr needs to be installed.

```
data %>%
     summarise(age= mean(age, na.rm = TRUE),
               bili = mean(bili, na.rm = TRUE),
               GOT = mean(GOT, na.rm = TRUE),
               GPT = mean(GPT, na.rm = TRUE),
               AP = mean(AP, na.rm = TRUE),
               GGT = mean(GGT, na.rm = TRUE)
               )
```

Figure 1: summarise function

The aforementioned function contains the following components. Mean in this case stands for the kind of summary that is desired from the specific column. na.rm = TRUE serves only as help to skip missing (NA) values.

The first line of code allows to string multiple operations together. By writing this code in R studio the following output is achieved.

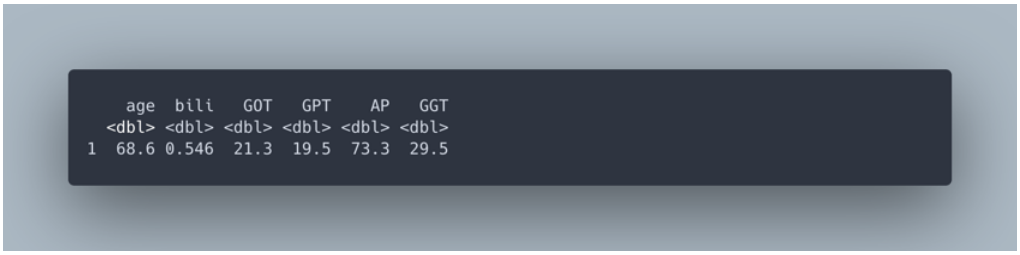


Figure 2: summarise function output

As it can be seen R calculates the average for some of the most important variables (age, bili, GOT, GPT, AP, GGT) of the dataset. For orientation, these values can be compared with the corresponding explanation in the list of abbreviations. From the results it can be seen that the three liver enzymes (GOT, GPT and GOT) have similar averages ranging from 19,5 (GPT) to 29,5 (GGT).

4.2 Data Analysis

First of all,when the built-in functions are not sufficient to accomplish a task, external packages need to be installed and later on loaded as libraries. These packages contain special functions and features that are not part of the standard version of R. In this specific case it is about the randomForest and the caret package. The first one is needed to create the random forest model while the second one is crucial to create a so-called confusion matrix. A confusion matrix is a tool to visualize the output of certain models like regression or classification models.

At this point, the desired dataset can be imported into R as this is the fundament for the whole evaluation. As mentioned before, data preparation is also a key step in the process because thereby errors can be cut down to the minimum and a more efficient analysis is enabled. In this case this means that the dataset needs to contain all the relevant information for the desired target. Also, certain columns that are not related to the experiment can be omitted.

After the data importation, the empty cells need to be handled. Having missing values in a medical dataset is quite normal since not every patient went through all the

possible tests and examinations. To run an efficient analysis, these empty cells must be omitted or deleted. In other words, there is a function that deletes rows where some cells are empty to provide a more compact set of data.

As the next step the values of which the prediction probability is searched needs to be formatted as a factor. A factor in R studio means that it is a categorial variable. Which means as much as that this column has a fixed selection of possibilities. To convert the desired category into a factor, the factor function is used. The seed can now be set, which is very meaningful to the experiment because it starts the random generator at a specific point. This serves for the reproducibility of the analysis as when the script is run again the same random numbers are used.

Now it comes to a very important step. Namely the breakdown of the data into a testing set and a training set. With use of the sample function the data is randomly split between the test set and the train set. The distribution is usually determined at 80% to 20% or 70% to 30%, with the larger percentage of the data being assigned to the training set. As mentioned in the literature review, the random forest model is a supervised machine learning algorithm, which means that as it is already in the name, the machine needs to train with the data in order to test the learned patterns on the basis of the testing set.

The random forest can now be created with help of the random forest function which was initially made possible by the installed the package of the same name. The function consists of a few components, namely the target variable (factor), the required dataset (train data) and the desired number of decision tress (ntree) for the model.

As the next step, the predict function is used to predict the preferred values based on the input data and check the outcome with help of the test set.

To visualize everything the previously addressed confusion matrix comes into operation.

```
library(caret)
library(randomForest)

View(data)

clean_data <- na.omit(data)

clean_data$color <- factor(clean_data$color)

set.seed(123)

trainIndex <- sample(1:nrow(clean_data), 0.8*nrow(clean_data))

trainData <- clean_data[trainIndex,]

testData <- clean_data[-trainIndex,]

rfModel <- randomForest(color ~ ., data = trainData, ntree = 500)

predictions <- predict(rfModel, testData)

confusionMatrix(predictions, testData$color)
```

Figure 3: code for the random forest model

This is the code used so far in R. All mentioned functions are visible here as well as the loading of the additional packages. The confusion matrix then visualizes the outcome of the applied random forest model.

In this case the input dataset consists of 11 variables and 3151 observations. The containing variables are as follows: gender, age, nationality, bili, GOT, GPT, AP, GGT, urea, urine density and urine color.

The goal is to get the accuracy with which the urine color can be predicted with use of the mentioned variables. By the way of explanation this means that if all variables except for urine color are given the accuracy with which it can be predicted is the desired outcome of this model.

In the confusion matrix there a several interesting key figures to note. The most important obviously the mentioned accuracy. But along with the accuracy there are also other results as well, such as 95% confidence interval, no information rate, p-value and Kappa.

Additionally, there are also the statistics by class. In this case the factor urine color has 8 possibilities: regular, water bright, yellow, light yellow, dark yellow, light brown, red brown and clouded.

The 95% confidence interval gives two values. With a certainty of 95% there can be said that the true average lies between these two values for the model.

The no information rate also has a percentage as output which must be surpassed by the accuracy in order to make the model significant. The percentage of the accuracy therefore needs to be higher than from the no information rate for the model to be considered significant.

The p-value is also related to the level of significance. Most of the time it is determined that a p-value lower than 0,05 means that the result is statistically significant.

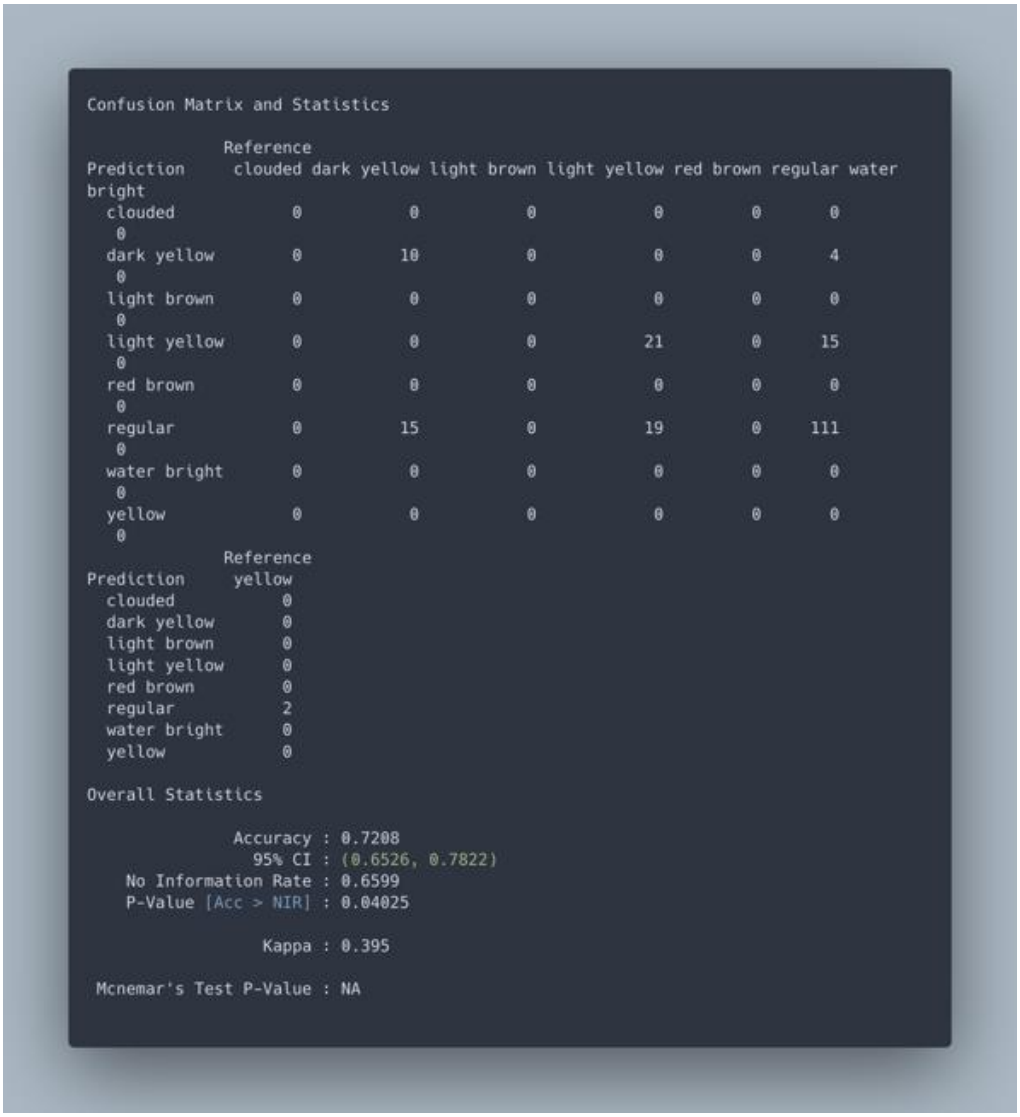
Kappa can have values between -1 and +1 and compares the overall accuracy with a random accuracy. The higher the Kappa value, the higher the accordance between the overall and random accuracy.

The results of the random forest model as well as the statistics by class can be seen later in this chapter. Important to not is that according to several sources a good accuracy for a random forest model lies somewhere between 70% and 90%. Anything higher than 90% therefore is considered outstanding.

With use of the head function the first few rows of the used data set can be displayed in R studio to get an overview. To get an idea of what this looks like it can be found in the Appendix chapter at the end of the thesis.

Figure 4: head function

The figure attached to the Appendix gives a better understanding of how the used dataset looks like. The different values as well as some empty cells can be seen.

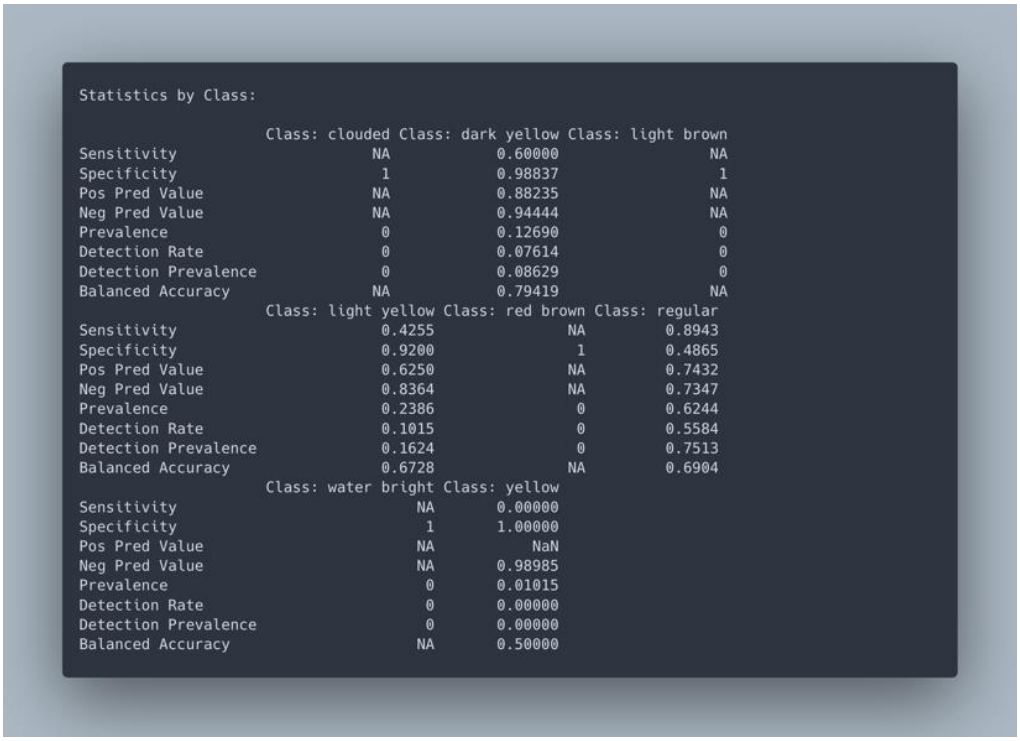


```
Confusion Matrix and Statistics

                Reference
Prediction       clouded dark yellow light brown light yellow red brown regular water
bright
  clouded              0           0           0            0         0       0
   0
  dark yellow          0          10           0            0         0       4
   0
  light brown          0           0           0            0         0       0
   0
  light yellow         0           0           0           21         0      15
   0
  red brown            0           0           0            0         0       0
   0
  regular              0          15           0           19         0     111
   0
  water bright         0           0           0            0         0       0
   0
  yellow               0           0           0            0         0       0
   0
                Reference
Prediction       yellow
  clouded             0
  dark yellow         0
  light brown         0
  light yellow        0
  red brown           0
  regular             2
  water bright        0
  yellow              0

Overall Statistics

               Accuracy : 0.7208
                 95% CI : (0.6526, 0.7822)
    No Information Rate : 0.6599
    P-Value [Acc > NIR] : 0.04025

                  Kappa : 0.395

 Mcnemar's Test P-Value : NA
```

Figure 5: results of random forest algorithm

As there can be seen in the figure above the accuracy with which the urine color can be predicted on the basis of the other 10 variables is a little over 72%. This is a pretty good percentage, especially when it is considered that with a sample size of roughly 3150 the dataset is not incredibly large. The 95% confidence interval verifies the accuracy as it states that the true average of accuracy lies somewhere between 65,26% and 78,22%.

The no information rate at 65,99% is also well below the accuracy (72,08%) which means that the results are statistically significant. This is also confirmed from the p-value of 0,04025. The Kappa indicates with a value of 0,395 a reasonably good match between the overall and the random accuracy.



Statistics by Class:

	Class: clouded	Class: dark yellow	Class: light brown
Sensitivity	NA	0.60000	NA
Specificity	1	0.98837	1
Pos Pred Value	NA	0.88235	NA
Neg Pred Value	NA	0.94444	NA
Prevalence	0	0.12690	0
Detection Rate	0	0.07614	0
Detection Prevalence	0	0.08629	0
Balanced Accuracy	NA	0.79419	NA

	Class: light yellow	Class: red brown	Class: regular
Sensitivity	0.4255	NA	0.8943
Specificity	0.9200	1	0.4865
Pos Pred Value	0.6250	NA	0.7432
Neg Pred Value	0.8364	NA	0.7347
Prevalence	0.2386	0	0.6244
Detection Rate	0.1015	0	0.5584
Detection Prevalence	0.1624	0	0.7513
Balanced Accuracy	0.6728	NA	0.6904

	Class: water bright	Class: yellow
Sensitivity	NA	0.00000
Specificity	1	1.00000
Pos Pred Value	NA	NaN
Neg Pred Value	NA	0.98985
Prevalence	0	0.01015
Detection Rate	0	0.00000
Detection Prevalence	0	0.00000
Balanced Accuracy	NA	0.50000

Figure 6: results of random forest algorithm by class

In Figure 6 the statistics by class are shown. By exploring statistics by class, insights into the random forest model for each class individually are accessed. This information can be beneficial to better understand the behavior of the model for different classes. Sensitivity, Specificity, Pos Pred Value, Neg Pred Value, Prevalence, Detection Rate, Detection Prevalence and Balanced Accuracy are all key figures that can be derived from these statistics. The above values are briefly explained.

Sensitivity indicates the proportion of successful cases that are considered correct by the random forest model among all cases, positive ones as well as negative ones. It displays the suitability of the model to recognize positive cases and is therefore also called the true positive rate.

Specificity in contrast to sensitivity, indicates the proportion of negative cases that were correctly detected by the model. This value is also called true negative rate. Both positive and negative cases are also included for specificity.

Pos. Pred. Value indicates the proportion of positively identified cases among the positively classified cases. Pred stands for prediction for both, pos pred value and neg pred value.

Neg. Pred. Value indicates the proportion of negatively identified cases among the negatively classified cases. Thus, both the pos pred value and the neg pred value are

very similar in terms of sensitivity and specificity with the only difference that that either only positive or only negative cases are considered

Prevalence refers purely to the proportion of correct positive cases and provides information on the distribution of positive cases. To get to this variable the correctly detected cases are divided by the total cases.

Detection Rate is quite similar to sensitivity. The difference is that the detection rate shows the proportion of positive detected cases without taking the false negatives into account.

Detection Prevalence indicates the number of positively predicted cases divided by the total predicted cases. Contrary to prevalence, true positive and true negative cases are considered here as well.

Balanced Accuracy refers to the average of sensitivity and specificity. Similar to total accuracy, over 70% is considered good and over 90% is considered outstanding.

The explanations show that the best results were obtained for the categories dark yellow, light yellow and regular.

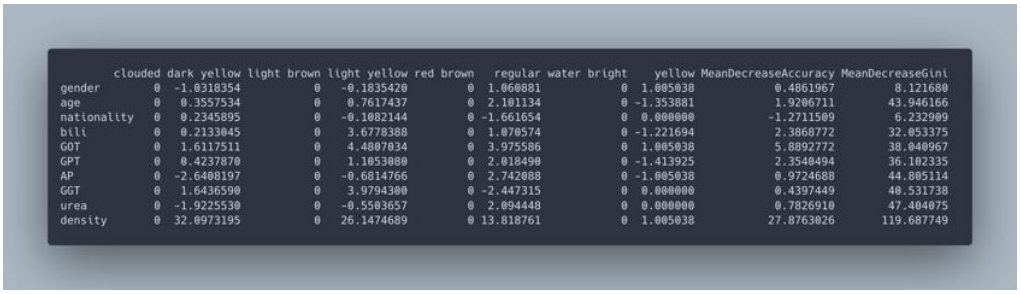
The category dark yellow especially has particularly superior results in almost all measurement values. Specificity (0,98837), pos pred value (0,88235) and neg pred value (0,94444) are very good, but also the balanced accuracy is impeccable with 79,419%.

4.3 Further Steps in the Evaluation

The code used so far in the analysis can be extended in several ways. One possibility is to compute the importance of the variables used in the random forest model. That means that among the variables used in the model, every single value can be assigned an importance as to how important the respective variable is in term of the predictive accuracy.

Figure 7: importance function

With the help of the importance function this operation can be performed. Here the results are shown of how this looks like for the conducted analyses when the function displayed is applied.



	clouded	dark yellow	light brown	light yellow	red brown	regular water	bright	yellow	MeanDecreaseAccuracy	MeanDecreaseGini
gender	0	-1.0318354	0	-0.1835420	0	1.060881	0	1.005038	0.4861967	8.121680
age	0	0.3557534	0	0.7617437	0	2.101134	0	-1.353881	1.9206711	43.946166
nationality	0	0.2345895	0	-0.1082144	0	-1.661654	0	0.000000	-1.2711509	6.232909
bili	0	0.2133045	0	3.6778388	0	1.070574	0	-1.221694	2.3868772	32.053375
GOT	0	1.6117511	0	4.4807034	0	3.975586	0	1.005038	5.8892772	38.040967
GPT	0	0.4237870	0	1.1053080	0	2.018490	0	-1.413925	2.3540494	36.102335
AP	0	-2.6408197	0	-0.6814766	0	2.742088	0	-1.005038	0.9724688	44.805114
GGT	0	1.6436590	0	3.9794300	0	-2.447315	0	0.000000	0.4397449	40.531738
urea	0	-1.9225530	0	-0.5503657	0	2.094448	0	0.000000	0.7826910	47.404075
density	0	32.0973195	0	26.1474689	0	13.818761	0	1.005038	27.8763026	119.687749

Figure 8: output of importance function

In this matrix all variables are assigned values that show the importance and that even for the individual output possibilities. It is quite clear that the value density, which is listed last, is by far the most important for the predicted accuracy. This becomes visible as the highest value results for some output possibilities.

4.4 Use of the Data Analysis for the Medical Facility

The analysis shows the accuracy with which the urine color of patients can be predicted based on the values gained through a small hemogram. This analysis has the best benefit for the company. Important conclusions as well as reverse conclusions can be drawn from it. The greater focus, however, lies on the reverse conclusions. The analysis can predict quite accurately which color the patient's urine should have based on the blood and liver values. Basic blood and liver tests as well as basic urine tests are conducted whenever a new patient comes to the medical institution for treatment. With the help of the values found out the company can compare the actual color of the urine with the predicted color. If the two colors do not match the probability is high that either the blood values or the liver values are not in order. The basic tests at the start of the treatment are a small hemogram which contains the blood and liver values used in the analysis. More specific tests are then conducted during the stay. The analyses can be used to determine if a patient should go through all additional tests or if additional tests would make sense at all. These additional tests are more complex and time-consuming for the company and thereby also more expensive. Several examinations are available for selection here. First in line is the big hemogram which is not conducted at the beginning and contains much more blood and liver values. But also tests and examinations that are much more complex

can be carried out or saved by using this method of data prediction and the resulting comparison. This concerns inspections such as computed tomography (CT), magnetic resonance imaging (MRI) and ultrasound examination, all with the aim to investigate the liver more accurately.

It can therefore be said that the company can optimize internal processes and procedures by the help of the conducted analysis. This has not only the early detection of anomalies as a consequence but also can help the company to save money on tests that are not necessary in certain situations.

On top of that the usage of data prediction can have even more internal as well as external advantages for the company. An internal factor could be the start of the inclusion of multiple data related analyses. Through the confirmation that processes like this actually work representatives of the company might be more willing to carry out many more analyses to optimize processes in the firm.

A positive external aspect is of course that through processes like this the institution can stand out of the competition which do not apply such analyses. Also, patients might feel more comfortable as a result because of the implied modern and reliable methods.

4.5 Recommendations for the Medical Facility

Since the experiment carried out shows success the most essential of all recommendations is to start analyzing more and more data in order to draw even more knowledge out of it. One possibility would be to hire a dedicated employee who, in addition to analyzing the medical data, generally monitors and analyzes all internal data. This goes hand in hand with the collection of data. Having more data accessible is by all means advisable for the company. Also, the comparison with data that is even older would be useful to see if certain things change with the turn of time. Since the company exists for more than hundred years a digitization of older analog data would be helpful as well. Through that more data could be included to the analyses which would enable even more accurate predictions. Changes that are perhaps generational could also be determined in this way.

Furthermore, it would be a great advantage to advertise with such modern methods and procedures in order to stand out from the competition. In this way, the institution could set a certain example to other companies by testing and introducing new

methods in addition to the established ones. This would reinforce that more companies start to analyze their own data and gain knowledge from it to help themselves in their development.

4.6 Further Research

In future research and related data analyses there are several optimizations that can be made. Of course, a larger dataset would be helpful in cases like this. Especially when it comes to machine learning and data prediction a larger dataset gives more opportunities to understand certain patterns and trends in the data provided.

Another crucial aspect would be to have similar sized datasets which contain the same tested values from different medical institutions. Trough that the results and accuracies could be compared and can be understood even better. Additional advantages can arise from comparing the results to medical facilities not only in the same country but with establishments from around the globe.

In order to avoid empty cells and therefore the data cleaning process which cuts the dataset down by a little it would be beneficial to have datasets from companies or hospitals that investigate all their patients in all medical values. More accurate processing of the data would of course also help in future research. This means that some data was incorrectly fed into the relevant system during recording. Often this has consequences that only become visible during the first attempts at analysis. Irregularities in the different columns of the dataset are meant. These irregularities can occur in the form of mixed characters for example. For instance, if in a specific column sometimes periods and sometimes commas are used indicate the decimal point of a number.

After consultation with the company, a few more goals for future research could be set. It would be best for the institution if the urine test alone could be used to draw conclusions about the other values. Urine tests are not only the least expensive but also by far the fastest to conduct and to evaluate. In view of this fact, it is comprehensible that successful methods that could help with this problem are desired by the medical company.

5 Conclusion

Through the conducted analysis there can be seen that even on a small scale, methods like data prediction can certainly show success. These methods have the potential to revolutionize the way in which medical companies operate regarding the collected data but also to the handling of it as a whole. As a result, a possibility emerges how the medical institution can optimize internal processes and therefore is able to save on costs. A method that saves costs in a company has of course many advantages. The money saved can be used for other purposes, such as further research with data, or simply to increase the company's profits in order to shed light on the economic side as well. Not only economic advantages can be achieved, but also treatments can be improved. As Banerjee et al. (2017) stated, both the patient and the company can benefit from such processes. Because of the fact that the patient benefits from it naturally, also brings an economic advantage that can be seen as a unique selling point. When patients experience positive outcomes, and improved well-being, it not only benefits their personal health but also leads to economic advantages for all stakeholders within the company.

This thesis can also be seen as an incentive to do even more research with medical data. Of course, such studies are even more significant on a larger scale, but as long as significant results can be achieved, any kind of data analysis is important, no matter on which scale. Even when examining data on a smaller scale, the potential for meaningful results should not be underestimated. Ultimately, the goal of the thesis was reached because it could show what possibilities can arise for a medical company through data analysis, respectively through data prediction and supervised machine learning algorithms to be precise. As data predictions continues to advance it is likely that it leads to a more efficient and effective health care system in general.

List of Abbreviations

bili: Bilirubin. Bilirubin is a red-orange breakdown product of the red blood pigment and thus a bile pigment. It is a waste product that is handled in the liver and drawn out by the bile. As reported by the Miller-Keane Encyclopedia (2003) laboratory tests in contemplation of measuring the level of bilirubin in the blood are crucial in the identification process of liver dysfunction and in the judgement of hemolytic anemias (a disease where red blood cells are faster destroyed then the human body can reproduce them). Bilirubin plays an important role in the human body and is responsible for several processes. An increased value, however, can indicate various health problems. For example, diseases such as hepatitis, biliary obstruction or inflammation in the bile ducts can be the trigger. A blood test determines total bilirubin, but also its conjugated and unconjugated forms.

Unconjugated bilirubin which is also known as indirect bilirubin is not soluble in water and is attached to albumin in the brood circulation, conjugated bilirubin on the other side (direct bilirubin) is soluble in water and has already been processed by the liver (Miller-Keane Encyclopedia, 2003).

According to Fevery (2008) bilirubin is an endogenous compound that can be poisonous under particular circumstances, on the contrary the author states that a light and unconjugated hyperbilirubinanmia might help in order to prevent tumor development and preserves against cardiovascular diseases. Furthermore Fevery (2008) explains that bilirubin counts in serum are frequently increased under various clinical conditions. These are studied in detail and the driving factors are described.

GOT: Glutamic oxaloacetic transaminase. Glutamic oxaloacetic transaminase is a liver enzyme that plays a major role regarding the metabolism of amino acids. It is mostly found in various body tissues and in serum, but especially in the heart and liver (Miller-Keane Encyclopedia, 2003). Additionally, there is stated that the enzyme is released into the serum when human tissue (mostly in the liver) is injured, which again primarily applies to the liver and heart. Conversely, an elevated concentration of glutamic oxaloacetic transaminase in the body may therefore indicate liver or heart disorders or diseases. The glutamic oxaloacetic transaminase enzyme therefore is an important indicator for the function of the liver. Diseases such as cirrhosis of the liver, fatty liver or hepatitis, for example, are diseases and conditions that can increase the

glutamic oxaloacetic transaminase level in the blood. It has to be mentioned that increased glutamic oxaloacetic transaminase values can also have a different origin as well. Examples of this are recent heart attacks or recent heart surgeries as well as the regular use of certain medications or excessively exhausting physical activities. According to various sources, like the Farlex Partner Medical Dictionary (2012) the normal range for healthy people lies somewhere between 10 and 40 units per liter of blood. Depending on different tests and evaluations this range can slightly differ from laboratory to laboratory. In summary, the lowering of glutamic oxaloacetic transaminase in the blood usually means a healthy liver with respect to the factors listed above.

GPT: Glutamic pyruvic transaminase. Glutamic pyruvic transaminase which is also called alanine aminotransferase (ALT) is an enzyme that is most common found I liver cells. According to the Collins Dictionary of Medicine (2004, 2005) important information related to liver disease and heart disease can be drawn from the enzyme since glutamic pyruvic transaminase is released into the blood from damaged heart and liver cells. A high value can indicate many diseases. In particular, a high value is recorded in the case of obese people. Quite similar to glutamic oxaloacetic transaminase, the range of units from a healthy human being lies between 10 and 40 units per liter of blood. Increased values may be associated with liver disease or discomfort and further testing will be required to explain the elevated values. “Therefore, it is plausible to suggest that aminotransferases are surrogate biomarkers of "liver metabolic functioning" beyond the classical concept of liver cellular damage, as their enzymatic activity might reflect key aspects of the physiology and pathophysiology of the liver function” (Sookoian & Pirola, 2015). In general, glutamic pyruvic transaminase is known as the most important indicator of the function of the liver. The value can be lowered again if there are no drastic increases, for example by purification in the form of drinking water and tea in large quantities. Also, the following applies equally to glutamic oxaloacetic transaminase, high values refer to over 50 units per liter of blood for men and over 35 units per liter of blood for women.

GGT: Gamma-glutamyl transferase. Gamma-glutamyl transferase is a liver enzyme that plays an important role in the metabolic process of Glutathion. This substance is

one of the most important substances that plays a role in the detoxification process in the liver. This value is also included in any standard blood test, as it also provides information about the function of the liver. “Gamma-glutamyl transferase (GGT) is a second-generation enzymatic liver function test available for several decades, initially used as a sensitive indicator of alcohol ingestion, hepatic inflammation, fatty liver disease, and hepatitis“(Mason et al., 2010). Mason et al. (2010) further explain that Gamma-glutamyl transferase is considered an important indicator of the cardiovascular system and the value is also associated with chronic heart diseases in general. Neuman et al. (2020) determine in their work that there is a significant correlation between elevated gamma-glutamyl transferase levels and heart-related diseases such as atrial fibrillation, cardiac insufficiency or cardiac arrhythmias. Further they also state that there is a correlation to sleep apnea. During sleep, the respiratory tract narrows to such an extent that breathing becomes difficult and can even fail completely (Neumann et al., 2020). As with the other liver enzymes, however, an increased value can also be related to the intake of specific drugs or other factors. However, the value differs from the other liver enzymes in view of the range in healthy people. According to various sources the range for healthy men lies between 8 and 45 and between 5 and 35 for healthy women.

AP: Alkaline phosphatase. Alkaline phosphatase is an enzyme that occurs in many parts of the human body. Typically, in different forms of tissue. Most commonly found in the bones, liver and colon. It is very important for liver function and the associated metabolism. Siller & Whyte (2018) state that alkaline phosphatase is probably the most studied enzyme in all of medicine, an elevated level in the bloodstream can be seen as an indicator of skeleton disease or liver disease. Together with the previously mentioned liver enzymes, alkaline phosphatase forms the basis for accurate liver tests. Since in the experiment carried out, also some persons who are not of age are included, it is important to mention that the activity of alkaline phosphatase changes extremely during puberty (Zierk et al., 2017). This change leads to significant irregularities of gender specific and age specific origin. According to Sharma & Prasad (2014) the range alkaline phosphatase in serum of a healthy human being amounts to 20 up to 140 units per liter of blood, where it is to be considered that with adult persons half of this activity takes place in the bones. Remarkably high values can, for

example, also indicate blockages in the bile ducts, also it should be noted that measured values from children and pregnant persons are higher than normal. Like the three liver enzymes, alkaline phosphatase is also present in every small hemogram conducted.

CT: Computed tomography. Computed tomography is a diagnostic x-ray tool with many different clinical applications that captures cross-sectional images of the human body (So & Nicolaou, 2021). It is a modern and complex procedure that allows diagnostic imaging of the inside of the human body. According to Seeram (2018), radiology technologies must be able to maximize dose and image quality to provide excellent care for the patient.

The author further explains that several pictures from a specific body part or organ are taken from different angles and views to combine them all too. A so-called contrast medium is often injected before examination in order to make specific parts of the body particularly visible. Fractures, internal bleeding and different types of tumors can be detected and visualized in this way. Examinations can be carried out without any pain and without any risk for the patient. Edholm (1977) emphasizes that one of the great advantages of computer tomography is that function tests can be carried out on different organs, in which not only the individual organ but also the function of all parts of it, no matter how small, can be determined. Furthermore Edholm (1977) stated that different soft parts of the body are directly visualized and also the shape of internal organs can be determined precisely. A computed tomography scanner can be mostly found in public hospitals as well as radiological doctors' offices.

MRI: Magnetic resonance imaging. Just as computed tomography, magnetic resonance imaging is a diagnostic tool that makes the structure of organs and tissue visible. According to Grover et al. (2015) the evolution of magnetic resonance imaging (MRI) in the course of medical treatments has led to great advances in diagnostics. Especially that the exposure of the patient to harmful radiation can be prevented is important.

The authors continue to explain that with better availability and declining costs the usage of magnetic resonance imaging becomes more and more common in clinical

practice. By the use of magnetic fields, the humas body is scanned layer by layer. On this way the inside of the body can be made visible and harmful structural changes of organs and tissue can be detected. The biggest advantages of magnetic resonance imaging are their low intensiveness, the absence of radiation and of course the ability of multiplanar imaging (Essing et al., 2000). Today, this method is a crucial part of clinical practice, and it is impossible to imagine processes without it.

Bibliography

Armstrong, S. (2017). *Data, data everywhere: the challenges of personalized medicine*. BMJ: British Medical Journal, 359. https://www.jstor.org/stable/26951517

Baker, J. W., & Henderson, S. (2017). The Cyber Data Science Process. *The Cyber Defense Review*, *2*(2), 47–68. https://www.jstor.org/stable/26267343

Balakrishna, R., Anandan, R. (2020). Feature Classification and Analysis of Acute and Chronic Pancreatitis Using Supervised Machine Learning Algorithm. In: Solanki, V., Hoang, M., Lu, Z., Pattnaik, P. (eds) Intelligent Computing in Engineering. Advances in Intelligent Systems and Computing, vol 1125. Springer, Singapore. https://doi.org/10.1007/978-981-15-2780-7_28

Banerjee, A., Mathew, D., & Rouane, K. (2017). Using patient data for patients' benefit. *BMJ: British Medical Journal*, *358*. https://www.jstor.org/stable/26941798

bilirubin. (n.d.) *Miller-Keane Encyclopedia and Dictionary of Medicine, Nursing, and Allied Health, Seventh Edition*. (2003). from https://medical-dictionary.thefreedictionary.com/Bilirubin

Blei, D. M., & Smyth, P. (2017). *Science and data science*. Proceedings of the National Academy of Sciences of the United States of America, 114(33), 8689–8692. https://www.jstor.org/stable/26487026

Borgheresi, R., Barucci, A., Colantonio, S. *et al.* (2022). NAVIGATOR: an Italian regional imaging biobank to promote precision medicine for oncologic patients. *Eur Radiol Exp* 6, 53 https://doi.org/10.1186/s41747-022-00306-9

Carter, D. and Sholler, D. (2016), Data science on the ground: Hype, criticism, and everyday work. *J Assn Inf Sci Tec*, 67: 2309-2319. https://doi.org/10.1002/asi.23563

Cetintemel, U. (2018). Predictive Analytics. In: Liu, L., Özsu, M.T. (eds) Encyclopedia of Database Systems. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-8265-9_80669

Columbus, L. (2017). *53% of companies adapting big data analytics*. Forbes. https://www.forbes.com/sites/louiscolumbus/2017/12/24/53-of-companies-are-adopting-big-data-analytics/?sh=5e0bb79139a1

Cote, C. (2021). 7 Data collection methods in business analytics. Harvard business school online https://online.hbs.edu/blog/post/data-collection-methods

Edholm, P. (1977). Computed Tomography - A New Technique in Diagnostic Radiology. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, *195*(1119), 277–279. http://www.jstor.org/stable/77186

Essig, M., Schoenberg, S. O., Schlemmer, H. P., Metzner, R., & van Kaick, G. (2000). Funktionelle Magnetresonanztomographie in der Neuroradiologie [Functional magnetic resonance tomography in neuroradiology]. *Der Radiologe*, *40*(10), 849–857. https://doi.org/10.1007/s001170050843

Fevery J. (2008). Bilirubin in clinical practice: a review. *Liver international : official journal of the International Association for the Study of the Liver*, *28*(5), 592–605. https://doi.org/10.1111/j.1478-3231.2008.01716.x

Floridi, L., & Taddeo, M. (2016). Introduction: What is data ethics? *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, *374*(2083), 1–5. http://www.jstor.org/stable/26115816

Glutamic Pyruvic Transaminase. (n.d.) *Collins Dictionary of Medicine*. (2004, 2005). from https://medical-dictionary.thefreedictionary.com/glutamic+pyruvic+transaminase

glutamic-oxaloacetic transaminase. (n.d.) *Farlex Partner Medical Dictionary*. (2012). from https://medical-dictionary.thefreedictionary.com/glutamic-oxaloacetic+transaminase

glutamic-oxaloacetic transaminase. (n.d.) *Miller-Keane Encyclopedia and Dictionary of Medicine, Nursing, and Allied Health, Seventh Edition*. (2003). from https://medical-dictionary.thefreedictionary.com/glutamic-oxaloacetic+transaminase

Grover, V. P., Tognarelli, J. M., Crossey, M. M., Cox, I. J., Taylor-Robinson, S. D., & McPhail, M. J. (2015). Magnetic Resonance Imaging: Principles and Techniques: Lessons for Clinicians. *Journal of clinical and experimental hepatology*, *5*(3), 246–255. https://doi.org/10.1016/j.jceh.2015.08.001

Juavinett, A. (2020). Data Science. In *So You Want to Be a Neuroscientist?* (pp. 207–212). Columbia University Press. http://www.jstor.org/stable/10.7312/juav19088.28

Landgraf, G. (2018). Data Collection and Privacy. *American Libraries*, *49*(9/10), 14–15. https://www.jstor.org/stable/26494712

Liu, G., Rodgers, J., Milne, S., Rowland, M., McIntosh, B., Best, M., Lepinard, O., & Hanham, M. (2019). Machine-Learning Algorithms. In *Eyes on U: Opportunities, Challenges, and Limits of Remote Sensing for Monitoring Uranium Mining and Milling* (pp. 15–17). James Martin Center for Nonproliferation Studies (CNS). http://www.jstor.org/stable/resrep19699.6

Mason, J. E., Starke, R. D., & Van Kirk, J. E. (2010). Gamma-glutamyl transferase: a novel cardiovascular risk biomarker. *Preventive cardiology*, *13*(1), 36–41. https://doi.org/10.1111/j.1751-7141.2009.00054.x

Michard, F., & Teboul, J.L. (2019). Predictive analytics: beyond the buzz. *Ann. Intensive Care* 9, 46. https://doi.org/10.1186/s13613-019-0524-9

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *The Journal of Economic Perspectives*, *31*(2), 87–106. http://www.jstor.org/stable/44235000

Neuman, M. G., Malnick, S., & Chertin, L. (2020). Gamma glutamyl transferase - an underestimated marker for cardiovascular disease and the metabolic syndrome. *Journal of pharmacy & pharmaceutical sciences : a publication of the Canadian Society for Pharmaceutical Sciences, Societe canadienne des sciences pharmaceutiques*, *23*(1), 65–74. https://doi.org/10.18433/jpps30923

Parsons, D. (2017). *Demystifying evaluation: Practical approaches for researchers and users* (1st ed.). Bristol University Press. https://doi.org/10.2307/j.ctt1t89h20

Pirracchio, R., Petersen, M. L., Carone, M., Rigon, M. R., Chevret, S., & van der Laan, M. J. (2015). Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. *The Lancet. Respiratory medicine*, *3*(1), 42–52. https://doi.org/10.1016/S2213-2600(14)70239-5

Richterich, A. (2018). Big Data-Driven Health Surveillance. In *The Big Data Agenda: Data Ethics and Critical Data Studies* (Vol. 6, pp. 71–90). University of Westminster Press. https://doi.org/10.2307/j.ctv5vddsw.7

Ricker, B. (2018). Reflexivity, Positionality, and Rigor in the Context of Big Data Research. In J. Thatcher, J. Eckert, & A. Shears (Eds.), *Thinking Big Data in Geography: New Regimes, New Research* (pp. 70–88). University of Nebraska Press. https://doi.org/10.2307/j.ctt21h4z6m.9

Ross, P. T., & Bibler Zaidi, N. L. (2019). Limited by our limitations. Perspectives on medical education, 8(4), 261-264. https://link.springer.com/article/10.1007/s40037-019-00530-x

Saraswat, P. (2022). Supervised Machine Learning Algorithm: A Review of Classification Techniques. In: García Márquez, F.P. (eds) International Conference on Intelligent Emerging Methods of Artificial Intelligence & Cloud Computing. IEMAICLOUD 2021. Smart Innovation, Systems and Technologies, vol 273. Springer, Cham. https://doi.org/10.1007/978-3-030-92905-3_58

Seeram E. (2018). Computed Tomography: A Technical Review. *Radiologic technology*, *89*(3), 279CT–302CT.

Sharma, U., Pal, D., & Prasad, R. (2014). Alkaline phosphatase: an overview. *Indian journal of clinical biochemistry : IJCB*, *29*(3), 269–278. https://doi.org/10.1007/s12291-013-0408-y

Shu, X. (2020). DATA VISUALIZATION. In *Knowledge Discovery in the Social Sciences: A Data Mining Approach* (1st ed., pp. 70–90). University of California Press. https://doi.org/10.2307/j.ctvw1d683.6

Sieber, R., & Tenney, M. (2018). Smaller and Slower Data in an Era of Big Data. In J. Thatcher, J. Eckert, & A. Shears (Eds.), *Thinking Big Data in Geography: New Regimes, New Research* (pp. 41–69). University of Nebraska Press. https://doi.org/10.2307/j.ctt21h4z6m.8

Siller, A. F., & Whyte, M. P. (2018). Alkaline Phosphatase: Discovery and Naming of Our Favorite Enzyme. *Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research*, *33*(2), 362–364. https://doi.org/10.1002/jbmr.3225

So, A., & Nicolaou, S. (2021). Spectral Computed Tomography: Fundamental Principles and Recent Developments. *Korean journal of radiology*, *22*(1), 86–96. https://doi.org/10.3348/kjr.2020.0144

Sookoian, S., & Pirola, C. J. (2015). Liver enzymes, metabolomics and genome-wide association studies: from systems biology to the personalized medicine. *World journal of gastroenterology*, *21*(3), 711–725. https://doi.org/10.3748/wjg.v21.i3.711

Terkola, R., Antoñanzas, F., & Postma, M. (2017). *Economic evaluation of personalized medicine: a call for real-world data*. The European Journal of Health Economics, 18(9), 1065–1067. http://www.jstor.org/stable/45156949

Tsuji, Y. (2020). Medical Big Data in Japan. *Journal of Law & Cyber Warfare*, *8*(1), 153–168. https://www.jstor.org/stable/26915566

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 1-67. doi:10.18637/jss.v045.i03

Venezia, F. (2018). Identifying Indicators for Evaluation Data Collection. In J. McE. Davis & T. H. McKay (Eds.), *A Guide to Useful Evaluation of Language Programs* (pp. 27–34). Georgetown University Press. https://doi.org/10.2307/j.ctvvngs5.8

Vens, C. (2013). Random Forest. In: Dubitzky, W., Wolkenhauer, O., Cho, KH., Yokota, H. (eds) Encyclopedia of Systems Biology. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-9863-7_612

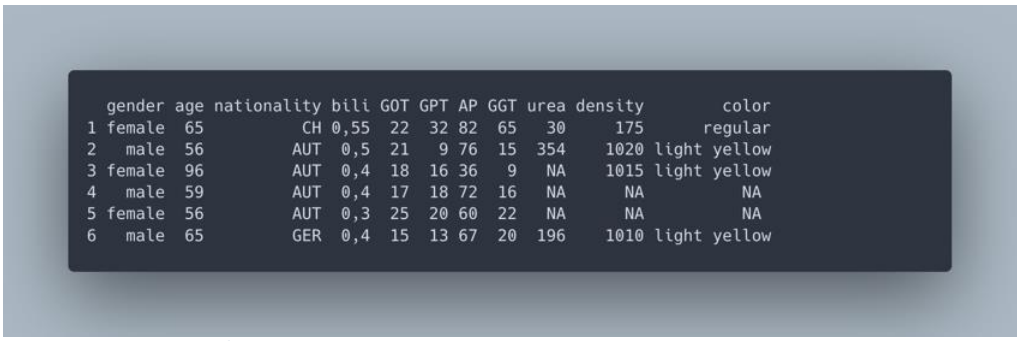
Zhang, L. (2018). Big data and medical research in China. *BMJ: British Medical Journal*, *360*. https://www.jstor.org/stable/26959342

Zhou, DX. (2015). Machine Learning Algorithms. In: Engquist, B. (eds) Encyclopedia of Applied and Computational Mathematics. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-70529-1_301

Zierk, J., Arzideh, F., Haeckel, R., Cario, H., Frühwald, M. C., Groß, H. J., Gscheidmeier, T., Hoffmann, R., Krebs, A., Lichtinghagen, R., Neumann, M., Ruf, H. G., Steigerwald, U., Streichert, T., Rascher, W., Metzler, M., & Rauh, M. (2017). Pediatric reference intervals for alkaline phosphatase. *Clinical chemistry and laboratory medicine*, *55*(1), 102–110. https://doi.org/10.1515/cclm-2016-0318

(2017). Random Forests. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7687-1_695

Appendix



	gender	age	nationality	bili	GOT	GPT	AP	GGT	urea	density	color
1	female	65	CH	0,55	22	32	82	65	30	175	regular
2	male	56	AUT	0,5	21	9	76	15	354	1020	light yellow
3	female	96	AUT	0,4	18	16	36	9	NA	1015	light yellow
4	male	59	AUT	0,4	17	18	72	16	NA	NA	NA
5	female	56	AUT	0,3	25	20	60	22	NA	NA	NA
6	male	65	GER	0,4	15	13	67	20	196	1010	light yellow

Figure 9: output head function