



University of Novi Sad

Faculty of Sciences

Department of Mathematics and Informatics



Student:

Ilija Doknić

Comparative analysis of applied classical statistical methods, models of machine learning and neural networks in the prediction of binary outcome

Master Thesis

Mentor:

Prof. dr Živko Bojović

Novi Sad, September 2024.

Acknowledgement:

First and foremost, I would like to express my deep gratitude to my mentor, Professor Živko Bojović. This work would not have been possible without his knowledge and dedication. Professor Bojović has greatly influenced not only my academic and professional career but has also given me invaluable life advice. What I am most grateful for is that Professor Bojović introduced me to a community of like-minded individuals who are passionate about technology and science.

One of those individuals is Professor Zoran Bukumirić, a professor of Medical Statistics at the University of Belgrade, who, along with Professor Bojović, made this thesis possible. In addition to providing the idea and data for this case study, Professor Bukumirić set an admirable example of how a true professional should conduct themselves. His disciplined and attentive approach made this work a pleasure to complete.

My gratitude also extends to Doc. Dr. Mirjana Mitrović and her team at the Clinic for Hematology at the University Clinical Center of Serbia. Over the years, they meticulously gathered the data required for this analysis, and Dr. Mitrović took the time to provide us with valuable advice.

While this thesis evolved into a medical case study, it initially started with a different focus. Professor Bojović and I, in collaboration with Marko Adžić, Global Tooling Head at Atos IT Solutions, aimed to develop conversational sentiment analysis for call-service evaluation. Although we had to abandon the project due to its infeasibility for a master's thesis, Mr. Adžić, with his vast business experience and acumen, taught me many valuable lessons. I want to thank him and express my hope to continue working on the project in the future.

I would also like to take this opportunity to thank Professor Nikola Obrenović. Although he was not directly involved in this project, his guidance and mentorship were of great importance during the time I was writing this thesis. Professor Obrenović shed light on the different paths I could take in my professional life and provided guidance that ultimately led me to the right place. Both Professors Obrenović and Bojović, as well as Professor Dušan Jakovetić, who organised the study programme, made me grateful that I chose to study Data Science in Novi Sad.

Finally, I want to thank my mother, Nada, and my brother, Matija, for helping with the final preparations of this document. I also extend my heartfelt gratitude to all my friends and family members. Their companionship and sacrifices, led by my father, Radojica, made my studies in Novi Sad possible. Among the family members already mentioned, I would like to specifically acknowledge my brothers, Vojin and Vukašin, my aunts, Gordana and Vera, my uncle, Slobodan, and my maternal grandparents, Radmila and Stojan. Lastly, I mention my paternal grandparents, Milosava and Miljan, who have passed away but will always remain in our hearts.

CONTENTS

1. Subject of research	1
2. Introduction to Data Science	2
2.1. Problems during model training	3
2.2. Data Processing.....	5
2.3. Evaluating models performance.....	8
2.4. Linear and Logistic Regression	13
2.5. Naive Bayes.....	15
2.6. Neural Networks.....	17
2.7. K-Nearest Neighbours.....	23
3. Data Science in Medicine.....	26
4. Related work.....	30
5. Research methodology	31
6. Results and discussion	41
7. Conclusion.....	45
References	46

LIST OF FIGURES

Figure 2.1 The desired line.....	3
Figure 2.2 Cross validation performed with 5 folds	5
Figure 2.3 The confusion matrix	9
Figure 2.4 An example of the ROC curve.	11
Figure 2.5 The changes in probability of a positive prediction as the ECOG_PS parameter changes	13
Figure 2.6 The representation of linear regression	14
Figure 2.7 The logistic curve i.e. sigmoid activation function	15
Figure 2.8 Samples of 10 librarians and 200 samples.....	16
Figure 2.9 The Typical Structure of a Neuron	18
Figure 2.10 The multiple layers that constitute a biological neural network.....	19
Figure 2.11 The architecture of the neural network	19
Figure 2.12 Image showing a computation of a value for a single neuron.	20
Figure 2.13 ReLU activation function	21
Figure 2.14 The architecture of a neural network with activations	22
Figure 2.15 A part of the neural network illustrating partial derivatives and backpropagation	23
Figure 2.16 K-NN classification	24
Figure 3.1 The surviving drafts of Nightingale's diagrams	27
Figure 3.2 Monthly mortality rates 1841-1849 in the hospital Semmelweis led.	28
Figure 3.3 Data visualisation of COVID-19. trends.....	29
Figure 5.1 The matrix of linear correlations	37
Figure 6.1 Logistic Regression ROC Curve.....	42
Figure 6.2 Changes in metrics as threshold changes.....	42
Figure 6.3 The summary plot of SHAP values	44
Figure 6.4 The waterfall plot of SHAP values for one example	44

LIST OF TABLES

Table 5.1 The summarisation of the entire dataset	35
Table 5.2 VIF values for attributes	38
Table 5.3 P-values for attributes.....	40
Table 6.1 Metrics for all the algorithms used.....	41
Table 6.2 Metrics for weighted regressions	43

1. Subject of research

The interdisciplinary field of data science aims to utilise ever-increasing quantities of data to draw insights and drive data-driven decision-making. It has the potential to revolutionise numerous sectors, including the healthcare industry. The field of data science is already being employed to analyse a range of patient data, including demographic information, basic health indicators, scans and other data points, with the aim of assessing the current state of disease or the likelihood of developing one. Furthermore, data science is employed at the macro level to combat epidemics and optimise the functioning of hospitals and entire healthcare systems.

This thesis explores the potential of data science methods for predicting venous thrombosis in patients with acute myeloid leukaemia. Thrombosis is a significant cause of morbidity and mortality among cancer patients, particularly those with leukaemia. Therefore, effective thrombosis prevention is a crucial aspect of cancer management. However, preventive measures against thrombosis may carry inherent risks and complications. Consequently, the application of thrombosis prevention should be limited to patients with a reasonable risk of developing thrombosis. Therefore, thrombosis prevention is carefully limited to patients with a reasonable risk of developing thrombosis, ensuring precision and thoughtfulness in the research.

In order to ascertain which patients are at risk, statistical and machine-learning algorithms will be employed to predict which patients with leukaemia will develop thrombosis. The data for this experiment was collected at the Clinic for Hematology at the University Clinical Center of Serbia. Researchers gathered data about individuals diagnosed with acute myeloid leukaemia (AML), including demographic information and various biomedical markers of interest. After a six-month follow-up period, they re-evaluated the patients to ascertain whether thrombosis had developed. This information about thrombosis represents the variable of interest, while the rest of the data is used to predict it retrospectively.

However, not all information collected is useful, and some may even impair performance. Therefore, we will examine which attributes are significant and what role they play in prediction. The algorithms applied include logistic regression, K-nearest neighbours, naive Bayes, and neural networks. The primary objective of this study was to identify the optimal model for this task and to evaluate and compare its performance.

2. Introduction to Data Science

The term "data science" is a relatively recent one, but it is already clear that it encompasses a great deal. As the name suggests, it is the science that studies how to extract knowledge from ever-larger quantities of data. It is a broad term that encompasses responsibilities of traditional roles such as statistician, data/business analyst, and database engineer, but also modern terms such as machine learning, cloud computing, big data, and artificial intelligence. The primary objectives of data science are to facilitate predictions, generate insights, inform data-driven decision-making, and develop artificial intelligence solutions. The foundation of data science is statistics, but the role of the statistician has evolved to encompass tasks that were previously the domain of other professionals. This shift is a consequence of the growing complexity of information technology, which has led to a need for a more diverse skill set in data-related work. While statistical expertise remains a crucial component, it was no longer sufficient for a statistician to possess only this knowledge. They must also demonstrate proficiency in areas such as programming, databases, economics, AI, and domain-specific fields.

In order to extract knowledge from data, a series of steps need to be taken. The initial step is to identify the data that is relevant to the problem at hand and to collect it. The method used to collect the data depends on the specific problem. The collection of data may be conducted manually through the implementation of surveys, observations, and measurements, or it may be automated through the deployment of sensors, web scraping, or the harvesting of data from internet users. The majority of data collected via the Internet and the Internet of Things (IoT) is automatically gathered, resulting in a significant increase in the overall volume of data. However, this data is predominantly unstructured, necessitating additional processing. While its quantity is considerable, the quality of this data is not always optimal. Manual labelling is essential to guaranteeing the highest quality of data, which is arguably the most crucial aspect of data science. Despite all advances in technology, manual labelling is still prevalent and represents a significant cost factor.

Secondly, the data must be transformed into a format that is suitable for analysis and provided with reliable infrastructure. This is particularly crucial when the data originates from disparate sources and formats. Combining, cleaning, and organising multiple sources into a single, consistent data set for storage, extract, transform, and load (ETL) methods are employed. The choice of storage depends on the type of data and its intended use. Storage types range from simple CSV files to SQL and document databases to data warehouses and data lakes. Once data has been collected, transformed, and loaded, the next step is analysis. Here, data scientists conduct an exploratory data analysis to examine patterns, ranges, and distributions of values within the data. Once a good view of the data has been obtained, the main part of the project is conducted, which is training a model.

Upon completion of the analysis, it is necessary to act on the results obtained. This typically entails disseminating the findings in the form of academic papers, reports, or visuals to relevant parties, whether they are technical or non-technical individuals. In the case of a data science project whose objective is the creation of a machine learning model, it is necessary to deploy the

model within another software and to monitor it in order to guarantee the continuous generation of high-quality outputs. Despite the fact that a data scientist must be aware of all of these processes, he is not typically acting alone. The collection of data is frequently conducted by domain experts, such as doctors and agronomists. The responsibility of storing data falls upon the data engineers, while the deployment of a model is the domain of the machine learning engineers.

2.1. Problems during model training

When training a data science algorithm, the goal is to utilise available data to obtain a mathematical model that will be capable of predicting the outcomes of future data. The process of making predictions with data that the model has never previously encountered is referred to as 'generalisation'. The objective is not to achieve the best results on the data that has already been seen. A model that has a low training error (the error on the known data) and a high generalisation error is said to be overfitting. This results in suboptimal models. This occurs because the algorithm learns the available data too well, word for word, and is unable to adapt to new situations. The image of an overfitted model is shown below.

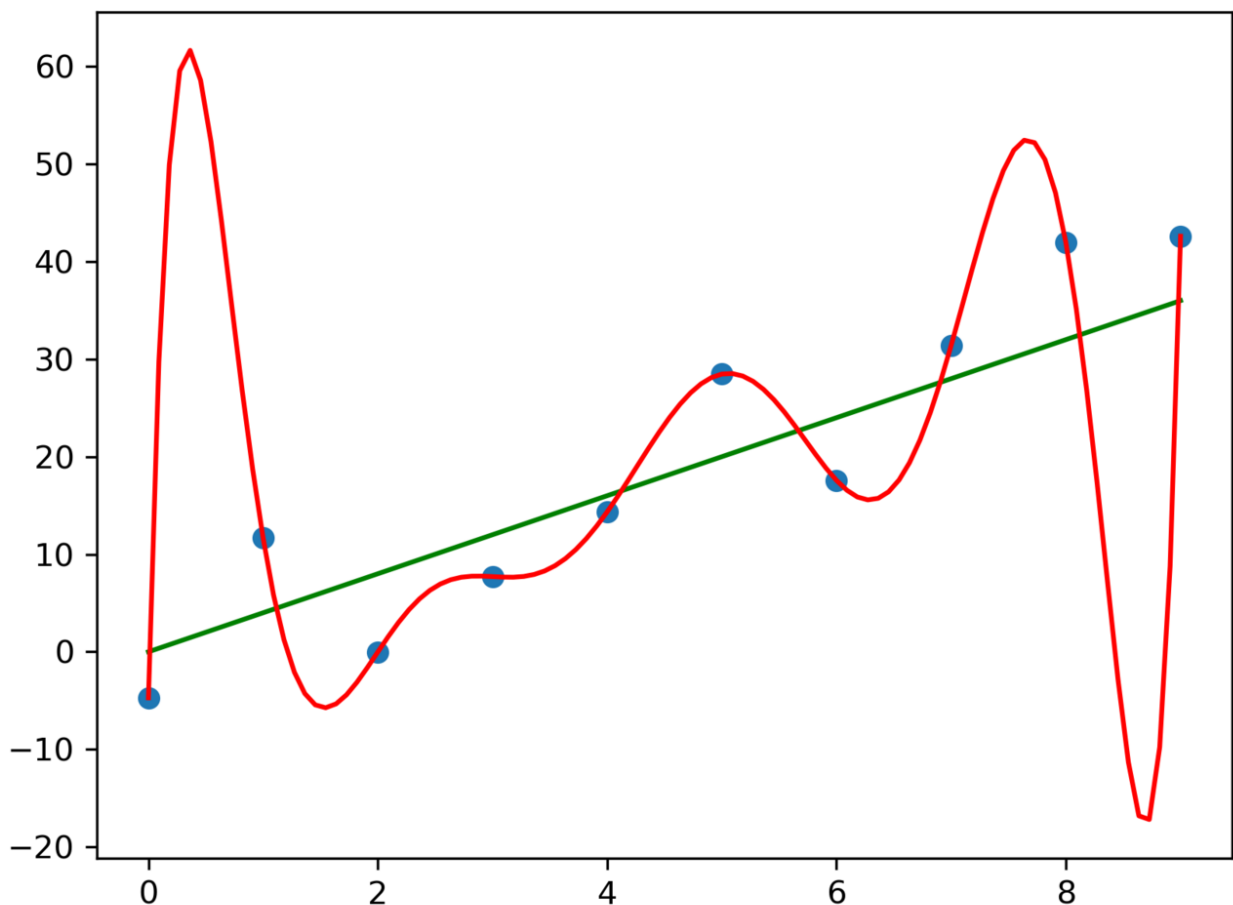


Figure 2.1 The image shows the desired line - green - and the overfitted line - red [32].

The opposite of overfitting is when a model is unable to identify patterns in data that have already been presented. This is referred to as underfitting, which is typified by elevated training and generalisation error. The primary cause of this phenomenon is the presence of an overly simplistic model or a deficit of sufficient parameters. Conversely, an excessively complex model or an abundance of attributes will result in overfitting. Therefore, it is of paramount importance to strike a balance between the model's complexity and the number of parameters, while also capturing the general trend of the data, in order to ensure the success of a data science project.

The phenomenon of underfitting is relatively straightforward to identify. The discrepancy between the dataset that was used for training and the desired outcome is greater than anticipated. However, it is not always straightforward to ascertain whether overfitting is occurring. One method for identifying this is to split the data into two distinct sets: a training set and a testing set. The model is then fitted on the training set, and the testing set is reserved for validation purposes only. This allows us to assess the model's performance in a real-world scenario, providing insights into the expected outcomes when applying the model in practice.

Nevertheless, if we were to repeatedly fit a model with different parameters and choose the final version based on the results on the testing set, the resulting data would be biased. The reason for this is that the model selected would likely have parameters specifically calibrated for that test set, which would produce overly optimistic results. This can be avoided by utilising an additional data set, referred to as a validation dataset. The validation dataset is employed for comparing the performance of different models that were fitted using a training set. Subsequently, the final model's performance is evaluated on a test set, thereby providing an unbiased assessment of its potential real-world performance.

The 60/20/20 method is the most commonly used approach to data splitting. This method allocates 60% of the data for training, 20% for validation, and 20% for testing. This is merely a heuristic that is applicable in general. Alternative data partitioning strategies are also frequently employed and should be tailored to the specific characteristics of the available data. In instances where the dataset is particularly large, exceeding 100,000 rows, a 98/1/1 split may be a viable option. The crucial factor is to ensure the inclusion of a high-quality test set that is both extensive and reflective of the underlying data distribution.

In cases where only a limited data set is available, cross-validation can be employed as an alternative approach that does not require the diversion of scarce data into a validation set. The cross-validation method involves the partitioning of the dataset into a specified number of subsets, or folds, referred to as K . At each stage of the process, one fold is set aside and the model is trained on the remaining $K - 1$ folds. This process is repeated K times, with the resulting values averaged to obtain the final estimate. Thus, there is no necessity to utilise a small validation set, which would inevitably yield biased outcomes. Furthermore, upon evaluating the performance of different models, it is possible to select the one that has demonstrated the most consistent success on average. Subsequently, with these parameters, the model can be retrained on the entire dataset. This approach avoids the loss of any data for validation, although the computational costs are higher, so this model is only applicable in cases where the dataset is

particularly small.

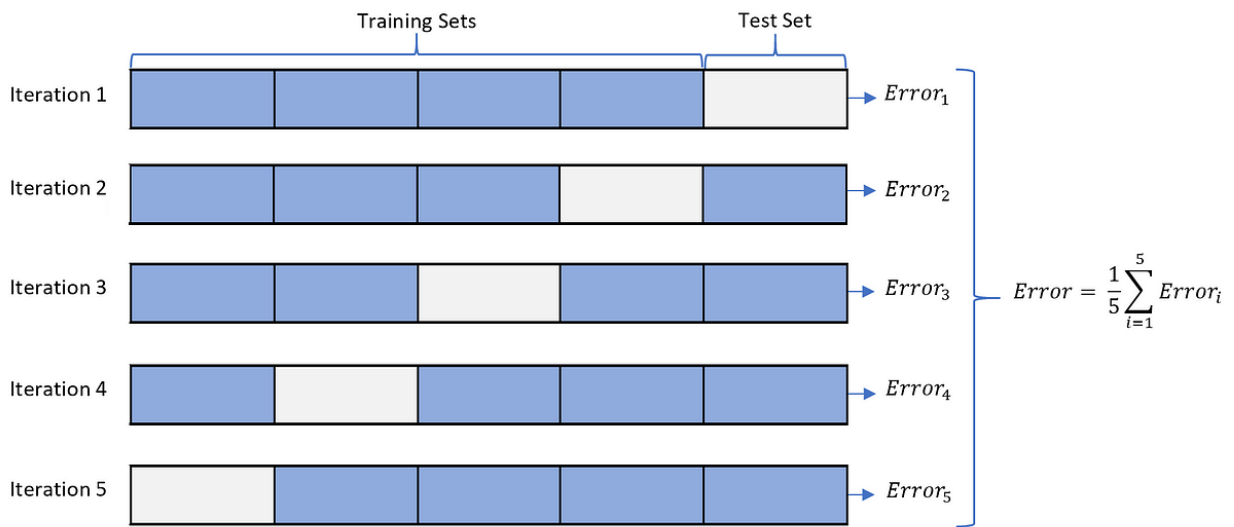


Figure 2.2 Cross validation performed with 5 folds [17].

Finding optimal parameters is a trial and error process. Models are constructed using a variety of parameters and then evaluated on the validation set, with the optimal parameters retained. A number of systematic approaches have been developed for this process, with two of the most commonly used being grid search and random search. In grid search, a number of values are specified for different parameters, and the model then attempts to find the combination that produces the best results. In contrast, random search requires distributions for each parameter being fitted, and at each iteration, the algorithm randomly samples values for each parameter and trains the model. In this approach, the number of iterations is defined.

2.2. Data Processing

As data science models become increasingly sophisticated, simple models have not lost their usefulness. Conversely, they are essential in numerous problems. In his book *Blink*, Malcolm Gladwell provides an example of the Count hospital in Chicago. The issue was that a considerable number of patients were presenting themselves at the emergency room with complaints of chest pain. Gladwell described the situation as follows: "From the beginning, the question of how to deal with heart attacks was front and centre. About 30 people a day came into the ER were worried that they were having a heart attack. And those thirty used more than their share of beds and nurses and doctors and stayed around a lot longer than other patients."

A change was required in the manner by which the hospital classified patients into urgent and non-urgent cases. A cardiologist named Lee Goldman proposed a straightforward criterion comprising just six parameters that would evaluate patients with chest pain and identify those at the highest risk of myocardial infarction. Goldman's criterion was largely derived from Bayesian logic and was evaluated over a two-year period, with the input of medical professionals. Goldman's rule demonstrated superior efficacy in two key aspects. It exhibited a 70%

improvement in correctly identifying patients who were not experiencing a myocardial infarction, while simultaneously demonstrating enhanced safety. The objective of a chest pain prediction model is to promptly direct patients with severe complications to the appropriate care units, namely the coronary and intermediate units. When left to their own devices, the doctors correctly identified the most serious patients between 75 and 89 percent of the time. In contrast, the algorithm correctly identified these patients in over 95 percent of cases.

Gladwell's term "thin slicing" which means filtering factors that matter from overwhelming numbers that don't is similar to the concept called curse of dimensionality in data science. It is often suboptimal to attempt to utilise all available information. Rather, it is more beneficial to focus on attributes that, in conjunction with other attributes, make a significant contribution to prediction. In relation to the use of superfluous information, Gladwell states: "Extra information is more than useless; it's harmful; it confuses the issues. What screws up doctors when they are trying to predict heart attacks is that they take too much information into account."

Strong mathematical evidence also supports the curse of dimensionality, demonstrating that as we transition from low to high-dimensional space, numerous issues emerge. As the number of parameters increases, the size of the parameter space grows exponentially, and a greater number of data points are required to populate it. In this high-dimensional space, every point lies on the edge, which renders distance metrics ineffective. Furthermore, Cover's theorem establishes that any partition of samples becomes linearly separable, leading to overfitting.

To counter the curse of dimensionality, sound feature engineering selection is necessary. Feature engineering can be defined as the process of constructing an optimal subset of features. It is a more sophisticated approach than feature selection, which merely entails identifying a subset. Feature engineering involves the construction of new and informative features from existing ones. One example of this is the use of correlated measures of height and weight to create a more informative feature known as the Body Mass Index.

Nevertheless, the selection of features represents a pivotal aspect of feature engineering. A variety of methodologies may be employed for this purpose, including:

- Univariate statistics
- Forward/backward feature selection
- Dimensionality reduction
- Algorithms with built-in feature selection
- Recursive feature elimination

There are numerous methodologies for the selection of parameters based on feature statistics. Running univariate linear regression and taking parameters with small p-values is a common way of selecting parameters. In addition to examining the relationship between predictors and targets, it is also important to consider the relationship between predictors themselves. The standard correlation coefficient is a straightforward approach that provides a broad overview. A more sophisticated method that offers a more detailed perspective is the variance inflation factor. This method predicts the values of one attribute based on the values of all the others.

Suppose the values for an attribute can be accurately predicted. In that case, it indicates that the attribute is an excess parameter and that there is multicollinearity, a detrimental phenomenon where attributes are strongly correlated with each other.

The forward feature selection process is initiated with the training of an algorithm on a single parameter. The algorithm's performance is then evaluated, and if it meets the requisite standards, the parameter is retained. The process is then repeated with the addition of a second parameter, and so on until either the performance ceases to improve or the specified number of parameters has been reached. In contrast, the backward feature selection process begins with the complete dataset and eliminates features that, when excluded, do not negatively impact performance.

The assumption underlying dimensionality reduction is that high-dimensional data can be represented by a lower-dimensional manifold. This is particularly the case when there is multicollinearity, since dimensionality reduction can both decorrelate data and reduce noise. The most commonly used dimensionality reduction algorithm is principal component analysis (PCA). This algorithm is based on singular value decomposition and has the goal of retaining axes that preserve the largest amount of linear variability. PCA allows the user to define the amount of variability that is to be retained or the number of features that are to be used to map the data.

In the context of empirical research, data scientists are likely to encounter instances where a value is absent, or "null." While some computer scientists have challenged the grammatical correctness of this term, given that "null" denotes the absence of a value, we will utilise this terminology for the sake of brevity. The term "null" signifies the absence of information regarding a specific value. It is crucial to ascertain the reason behind this absence. It could be that the value in question does not exist, such as the age of the eldest child for individuals who do not have children. Alternatively, the value may exist but has not been documented, such as a person's age.

In the event that the value is absent despite the knowledge that it is expected to be present, there may be a number of potential explanations for this phenomenon. Researchers have identified three distinct categories:

- Missing completely at random: This refers to instances where there is no discernible pattern or rationale behind the absence of data. To illustrate, some images lack an accompanying caption due to an oversight.
- Missing at random: parameters are deemed to be missing at random when the reason for their absence is related to the other parameters. At elevated temperatures, the sensors are unable to record specific data points, which results in the observed absence of these data points.
- Missing not at random: the reason for the parameter's absence is its own value. For example, individuals with either very low or very high income levels are less likely to disclose this information.

An understanding of the reasons for the absence of data enables more effective strategies to be employed in addressing the issue. But, there are limitations to what can be done. In instances

where the proportion of missing data in a column exceeds 5%, it is likely that the most appropriate course of action would be to exclude this column entirely. In instances where the number of null values is tolerable, one may employ a variety of techniques to fill the gaps. These include the use of rules, imputation methods based on statistics, or more complex statistical procedures. The application of rules is particularly useful when the reason for the absence of data is understood, and when the assumption of missing completely at random is valid. Imputation techniques, on the other hand, rely on the use of centrality measures, such as the mean, median, or mode, to fill the missing values.

Once the missing values have been removed, the subsequent step is to ensure that the features are scaled in a consistent manner. In order to achieve an optimal solution, algorithms that rely on calculating distances in the metric space, such as K-Nearest Neighbours or K-Means, require this step. In the event that the features in question possess different scales, the variables exhibiting higher values will exert a disproportionate influence on the models. The rationale behind this phenomenon can be understood by considering the intuitive understanding that, from a human perspective, 1 kg is equivalent to 1000 grams. However, this is not the case for an algorithm, as the distance of 1 is associated with one feature, while the other is associated with 1000, resulting in a significant discrepancy in influence. To circumvent this issue, it is essential to normalise or standardise all features, ensuring that they are within the range of 0 to 1 or adhere to a normal distribution with a mean of 0 and standard deviation of 1, respectively.

2.3. Evaluating models performance

One aspect that was not addressed in the previous chapter is the methodology for determining the optimal model and evaluating its performance. The optimal model is dependent on a multitude of factors, including the specific task at hand and the current state of the art. However, all models are evaluated using some form of performance indicators. The formulas used to assess model performance are referred to as metrics. There are a plethora of metrics employed for a vast array of tasks, including classification, regression, clustering, image segmentation, and machine translation. In this thesis, we will focus exclusively on metrics related to classification.

The most straightforward metric for classification is accuracy. In essence, accuracy is the ratio of correct predictions to the total number of predictions. It is a useful metric for comparing different models, as it provides a general indication of their performance. However, it is important to exercise caution when interpreting accuracy alone, as it can be misleading if the dataset is not balanced, i.e. if it contains a disproportionate number of positive examples. To illustrate this, consider an algorithm trained to predict a disease that affects one in a hundred people. A model that states that no one has the disease would have 99% accuracy, despite being completely useless.

In addition to not being able to work with unbalanced datasets the accuracy measure does not indicate the types of errors that are being made. In binary classification, there are four potential outcomes, two of which are correct and two of which are erroneous. An accurate outcome would be the prediction of the positive class for a positive example and the negative class for a negative

example. Mistakes, on the other hand, can be made in two ways. A negative example can be classified as positive, resulting in a false positive, and a positive example can be classified as negative, resulting in a false negative. This can be conceptualised as a judge making a decision in criminal court. The objective is to incarcerate a guilty individual and to release an innocent one. However, the judge could also incarcerate an innocent individual, resulting in a false positive, or refrain from sentencing a guilty person, resulting in a false negative.

Similarly, a data scientist must make a decision that is analogous to that of a judge, namely, which option to avoid. As the decision in law is dependent upon the time and context, so too is it in data science projects. In the context of disease detection, it would be preferable to inform a healthy patient that they may have a disease and should undergo further testing, rather than informing a sick patient that they are healthy. Conversely, in the context of spam prediction, it would be more beneficial to allow spam into the main folder than to mark an important email as spam.

The aforementioned metrics are typically represented in a tabular format, known as a confusion matrix, as illustrated in the accompanying image. The rows of the matrix correspond to the actual classes, whereas the columns represent the predicted classes. While alternative configurations of the confusion matrix are occasionally encountered in the literature, this particular implementation from the scikit-learn library is widely utilised due to its versatility in multi-class classification.

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Figure 2.3 The implementation of the confusion matrix in the Skit-learn library. The following acronyms are defined: TN (true negative), FP (false positive), FN (false negative), and TP (true positive) [23].

The confusion matrix also provides a foundation for deriving other metrics. Precision quantifies the proportion of positively predicted classes that are, in fact, positive. In scenarios such as spam detection or recommended systems, where it is crucial to ensure that recommended content is of real interest to consumers, high precision is a desirable outcome.

$$\text{Precision} = \frac{TP}{TP + FP}$$

The recall measure represents the percentage of true positive predictions among all positive cases. In situations where false negatives would be particularly costly, a high recall is to be preferred. One example of such a circumstance is fraud detection, where it is of paramount importance to identify malicious actors and to avoid the imposition of significant expenses. Conversely, flagging regular transactions as potentially fraudulent will result in them undergoing additional inspections, which is a less costly process. The following formula represents the recall measure:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Although there may be occasions when one is preferred over the other, it is almost always the case that both precision and recall should be within a reasonable range. Calculating the mean of the two values would not take this into account, since one value could be extremely high and the other extremely low, with the result being a mean that balances these out. In order to penalise models that have one extremely low value, we use harmonic means or F1 score.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Finding the right balance between false positives and false negatives is accomplished through the modification of the probability threshold utilised for making predictions. In most cases, the threshold is set to 0.5, which means that instances whose output is higher than that are classified as positive. Modifying this threshold results in alterations to the algorithmic process. An increase in the threshold will enhance precision, as the algorithm will require greater certainty for positive predictions. Conversely, a decrease in the threshold will lower the threshold for positive prediction, resulting in an increase in positive examples and higher recall.

A plot illustrating the impact of threshold alterations on a model is termed a receiver operating characteristic (ROC) curve. However, it is not plotted with precision and recall, but with analogous metrics of sensitivity and specificity. Sensitivity is simply another term for recall or true positive rate, while specificity is defined as follows:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

The ROC curve is illustrated in the following figure. The ROC curve has two axes. The horizontal axis is named the false positive rate, which is equal to 1 minus the specificity. The vertical axis is the true positive rate. The diagonal dotted line shows how a perfectly random model would act. The curve is obtained by varying the threshold and plotting the results. The further the curve is to the upper left corner, the better the overall performance of the algorithm. Thus, the area under the curve (AUC/C-score) is a logical measure of the algorithm's performance.

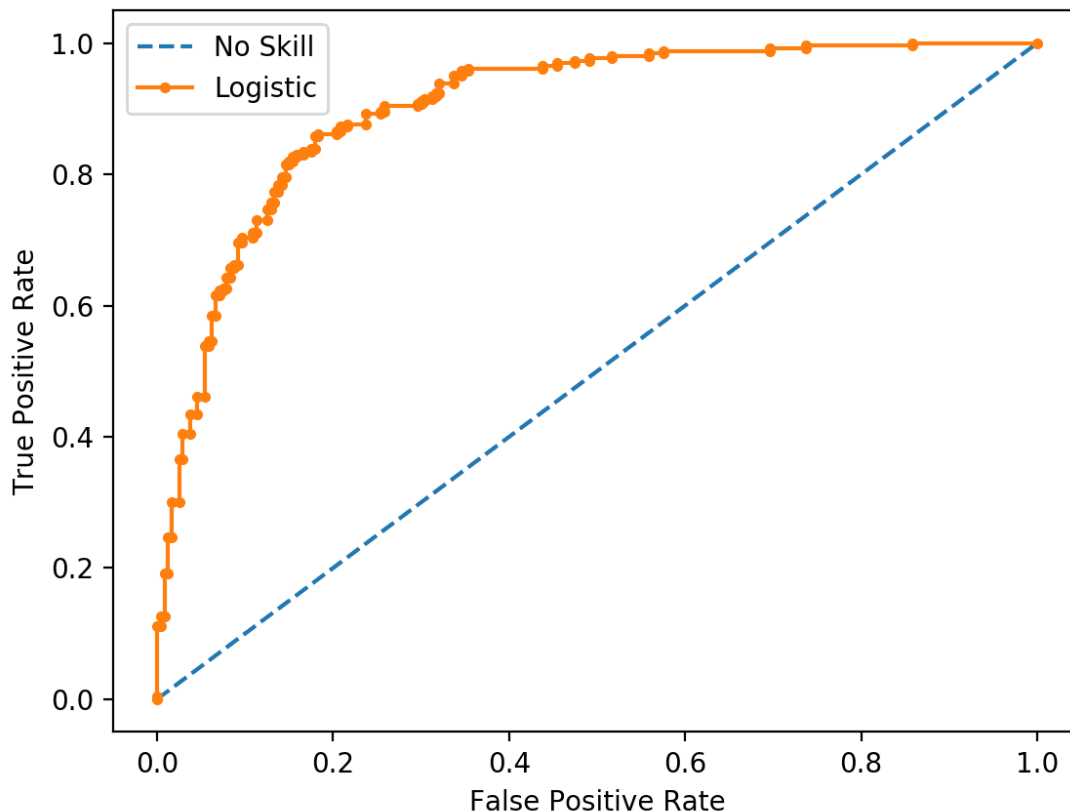


Figure 2.4 An example of the ROC curve. The orange line represents the performance of the logistic regression while the dotted line shows the performance of the random model [31].

Once different values have been assigned to the metrics in question, it becomes necessary to ascertain whether the resulting outcomes are favourable or otherwise. Does a 95% accuracy rate represent a high or low value for this metric? The answer is contingent upon the specific task at hand. In the context of the recognition of written digits, a well-known task for which numerous successful algorithms have been developed, a level of accuracy of 95% would be considered relatively low. Conversely, in the context of the prediction of stock market movements, a model with an accuracy of even 51% would be considered revolutionary and would potentially bring significant financial rewards to its creator. The optimal result is determined by an evaluation of the current state of affairs, human performance and the state of the art solutions.

The celebrated statistician John Turner once observed that: "All statistical models are wrong. But some of them are useful." If a model can be implemented in a system to enhance its functionality, it can be considered a useful model. A common objective of contemporary artificial intelligence models is to supplant human labour. In the development of models for tasks such as vision, language, logical reasoning, and so forth, where human performance is superior to that of machines, human-level performance serves as a valuable proxy for assessing tradeoffs of implementing an AI strategy.

Conversely, AI has been demonstrated to outperform humans in tasks that involve a multitude of variables, statistical reasoning, and scenarios where human bias may be a factor. These tasks encompass online advertising, content recommendation, transit time prediction, loan approval, and others. In such cases, it is prudent to utilise human performance as a benchmark rather than relying on existing state-of-the-art comparisons.

Thus far, we have discussed only intrinsic metrics; however, there are also extrinsic ones. Intrinsic focuses on intermediary objectives, while extrinsic focuses on evaluating performance on the final objective. For example, consider a spam-classification system. The ML metric will be precision and recall, while the business metric will be “the amount of time users spent on a spam email.” Intrinsic evaluation will focus on measuring the system performance using precision and recall. Extrinsic evaluation will focus on measuring the time a user wasted because a spam email went to their inbox or a genuine email went to their spam folder [24].

Nevertheless, the performance of the model is not the sole consideration when developing machine learning models. It is also important for models to be able to provide an explanation for the rationale behind specific decisions. Such insight enables researchers to understand the root cause of the problem more profoundly, thereby facilitating the implementation of effective solutions. In the event that a model indicates that a patient is at elevated risk of developing a disease, it would be highly beneficial if the model were to provide an explanation, thereby enabling the physician to prescribe an appropriate course of treatment. Similarly, individuals who are affected by the model must also be informed of the rationale behind specific decisions. In the context of a loan application, for instance, if a bank declines to provide financing, the applicant must be informed of the reasons for this decision and the conditions under which the bank would be willing to approve the loan.

A straightforward approach to determining the contribution of each parameter is to generate a partial dependence plot. These plots are constructed by taking a single row instance and repeatedly modifying one of its features while maintaining all others constant. Predictions are then made with the modified instance and stored. Finally, the predictions are plotted against different values of the modified parameter, resulting in a plot that illustrates how the prediction value changes as the parameter value changes. This provides valuable insight into the parameter's influence.

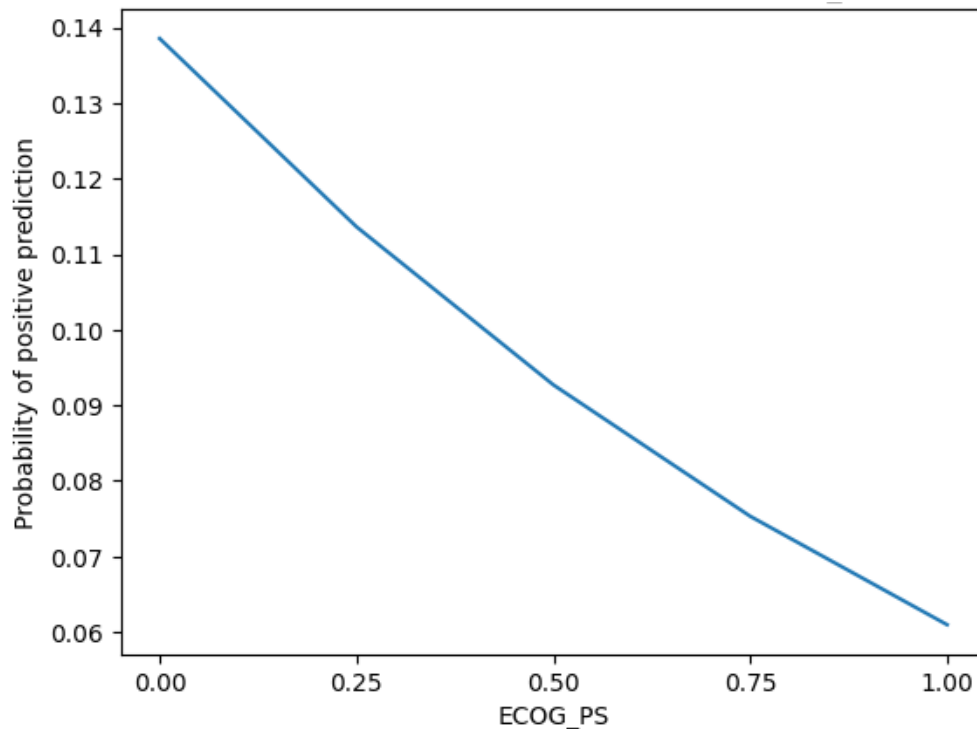


Figure 2.5 The graph illustrates changes in probability of a positive prediction as the ECOG_PS parameter changes

This approach however, carries a significant risk of yielding results that do not align with the broader data distribution patterns. An approach offering a slight remedy to this problem are SHAP values. The concept of Shapley values is inspired by game theory, and their primary purpose is to determine the contribution of each factor in a model, in order to identify those with the greatest impact. These values are calculated iteratively, taking into account all possible combinations of the factors of interest, as well as their order. For example, if we want to assess the importance of the following variables: 1) age, 2) gender, and 3) tumour size, we will analyse how the prediction made by the ML model changes when each of these variables is added or removed in combination with the other two. This process is repeated until all possible combinations of the presence/absence of variables and their order are considered. At the end of each combination, we obtain the Shapley value for a given variable, and the average of these values, calculated from all possible combinations, represents the final Shapley value for the variable in question [33].

2.4. Linear and Logistic Regression

The statistical methods of linear and logistic regression are widely used and form the foundation for more complex algorithms. The former is employed for regression tasks, whereby a numeric value (such as the price of a house) is predicted. In contrast, the latter is used for classification, whereby a predefined class is assigned to each instance (such as whether a patient is in the high-risk group for developing venous thrombosis). Both of these algorithms function by assigning a multiplier to each attribute, which is referred to as the weight. The linear regression formula, which also forms the basis of the logistic regression formula, is as follows:

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_nx_n$$

Where:

- y is the dependent variable (the outcome we're trying to predict)
- β_0 is the y-intercept (the value of y when $x = 0$)
- $\beta_1 \dots \beta_n$ are the slopes of the regression line (the change in y for a one-unit change in x)
- $x_1 \dots x_n$ are independent variables (attributes)

Moreover, the application of linear regression can be visually represented through the graphical illustration of a line that best fits the data set. The subsequent image presents an example of such a line. The blue dots represent data points, with the independent variable represented by the x-axis and the dependent variable represented by the y-axis.

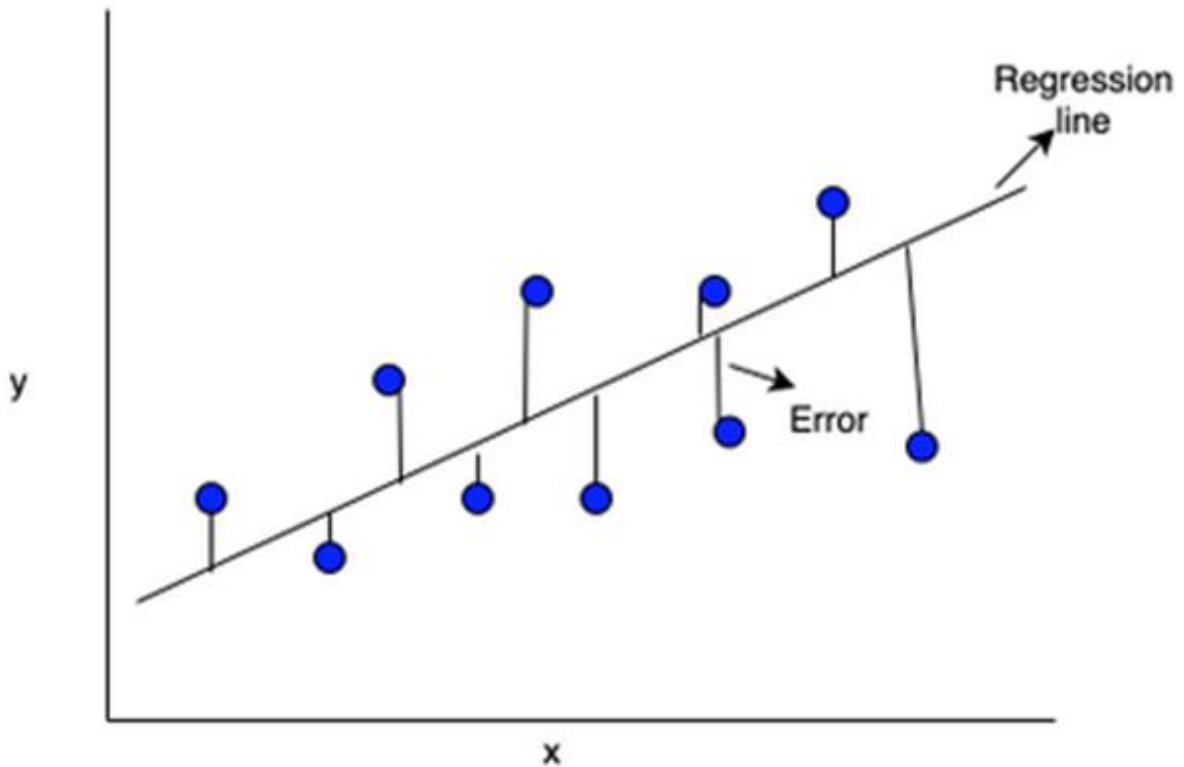


Figure 2.6 The representation of linear regression: the lines between the blue points and the regression line indicate the magnitude of the error for each point [23].

The distinction between the two lies in the manner of representation of the output variable, y . In logistic regression, this variable is compressed between 0 and 1, thereby indicating the probability of an instance belonging to the positive class. The act of squashing is performed using a sigmoid function, which is illustrated in the accompanying diagram. All negative values are mapped to a range between 0 and 0.5, which typically corresponds to the negative class. Values denoting the positive class are mapped to the range between 0.5 and 1.

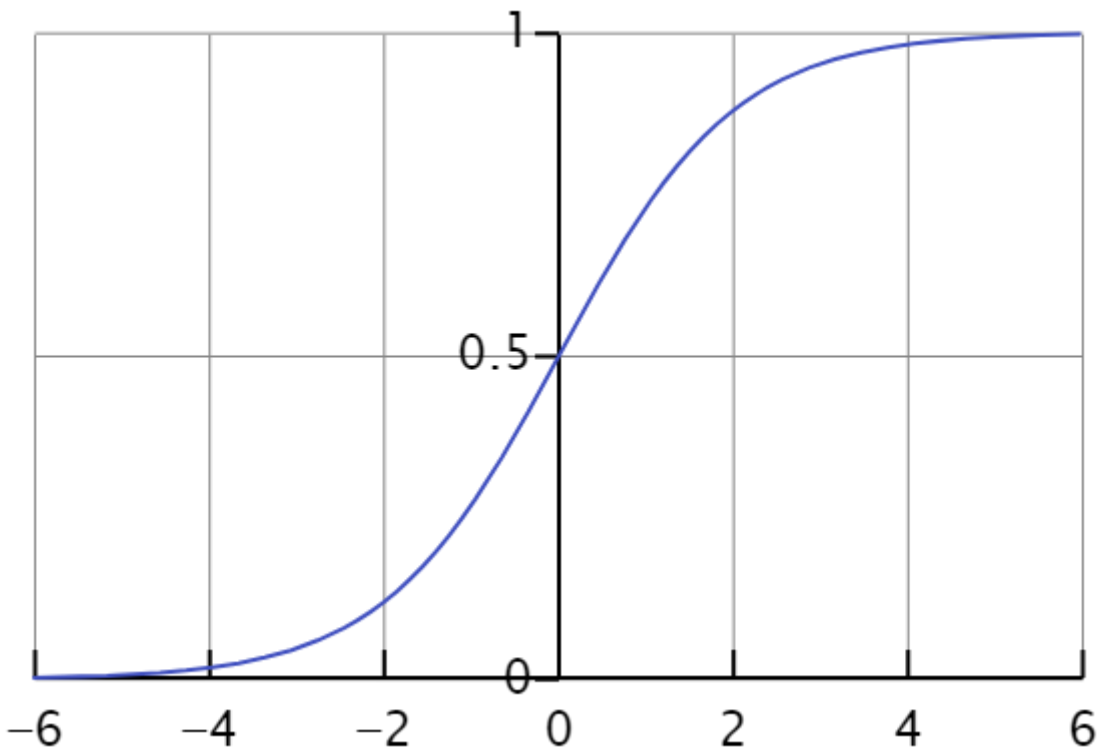


Figure 2.7 The logistic curve i.e. sigmoid activation function [30].

Linear regression is fit using least squares, and the quality of the fit is evaluated using RMSE and R-squared statistics. In logistic regression (unlike in linear regression), there is no closed-form solution, and the model must be fit using maximum likelihood estimation (MLE). Maximum likelihood estimation is a process that tries to find the model that is most likely to have produced the data we see. In the logistic regression equation, the response is not 0 or 1 but rather an estimate of the log odds that the response is 1. The MLE finds the solution such that the estimated log odds best describes the observed outcome. The mechanics of the algorithm involve a Quasi-Newton optimization that iterates between a scoring step (Fisher's scoring), based on the current parameters, and an update to the parameters to improve the fit [18].

2.5. Naive Bayes

The naive Bayes algorithm uses the probability of observing predictor values, given an outcome, to estimate what is really of interest: the probability of observing outcome $Y = i$ given a set of predictor values [18]. As the name suggests, the algorithm is based on the Bayes theorem, which gives probability that an event happens given that another event has already occurred.

One particularly illuminating example of Bayes theorem is the "Steve the librarian" puzzle, which was first presented by Daniel Kahneman in his book *Thinking, Fast and Slow* [19]. The problem can be summarised as follows: "Steve is very shy and withdrawn, invariably helpful but with very little interest in people or in the world of reality. A meek and tidy soul, he has a need for order

and structure, and a passion for detail.” Is Steve more likely to be a librarian or a farmer? The majority of individuals tend to assume that Steve is a librarian, as the description is more closely aligned with the characteristics typically associated with a librarian than with those of a farmer. However, this assumption is inaccurate.

What people fail to consider is a discrepancy in the number of male farmers and male librarians. In fact, there are at least 20 times more male farmers than there are male librarians. Even if all librarians were "meek and tidy souls" and the same were true of only one in fifteen farmers, there would still be more meek and tidy farmers than meek and tidy librarians. To illustrate this, we can take a representative sample of 200 farmers and 10 librarians (see figure below). For the purposes of this calculation, we can assume that the description fits 40% of librarians and 10% of farmers. When we perform the necessary calculations, we find that there are 4 librarians and 20 farmers in the sample, which means that the probability that Steve is a librarian is only 16.7%.

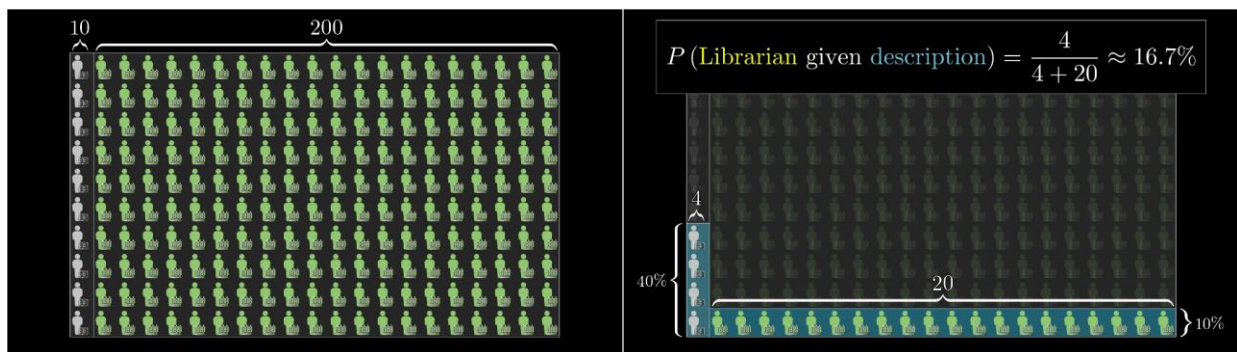


Figure 2.8 Left picture: Sample of 10 librarians and 200 samples; Right picture: Librarians and Farmers that fit description [20].

To put this into mathematical perspective we write the formula for Bayes theorem:

$$P(Y | X) = \frac{P(X | Y) \cdot P(Y)}{P(X)}$$

- $P(Y|X)$ is the value we are trying to predict and it is called posterior probability.
- $P(X|Y)$ is the likelihood or probability that X occurred given Y . This part is modelled from training data
- $P(Y)$ is the prior probability of the event happening meaning what are the chances of $Y = i$ if we don't know X
- $P(X)$ is a marginal probability

The marginal probability can be expanded using the law of total probability. In the case of a binary outcome, the formula can be written as follows.

$$P(Y = i | X) = \frac{P(Y = i) \cdot P(X | Y = i)}{P(Y = i) \cdot P(X | Y = i) + P(Y = j) \cdot P(X | Y = j)}$$

The Steve example featured a single parameter, X , which served as the basis for the description. In the real world, however, we are likely to encounter a multitude of parameters, each of which must be taken into account when making predictions. When X has multiple parameters, the formula is written as follows:

$$P(Y = i | X_1, X_2, \dots, X_p) = \frac{P(Y = i) \cdot P(X_1, X_2, \dots, X_p | Y = i)}{P(Y = 0) \cdot P(X_1, X_2, \dots, X_p | Y = 0) + P(Y = 1) \cdot P(X_1, X_2, \dots, X_p | Y = 1)}$$

In order to fit the model, one must first take a sample where $Y = 0$. This sample is then divided by the length of the entire dataset in order to obtain $P(Y = 0)$. The sample must then be divided by the count of all data in order to obtain the number of examples in the sample that have values X_1, X_2, \dots, X_p . Nevertheless, this approach becomes problematic when the number of parameters exceeds a few, as there are a multitude of combinations of X_1, X_2, \dots, X_p that lack matches in the training data. It is possible to envisage a model for the prediction of voting behaviour based on demographic variables. Even a sizable sample may not contain a single match for a new record, for example a male Hispanic with a high income from the US Mid-west who voted in the last election, did not vote in the prior election, has three daughters and one son, and is divorced. This is despite the fact that the model is based on just eight variables, which is a relatively small number for most classification problems.

In order to resolve this issue, it is necessary to make the assumption that the conditional probabilities are independent of one another. In other words, we assume that $P(X_j | Y = i)$ is independent of all other X_k for $k \neq j$. This allows us to separate this part of the formula, $P(X_1, X_2 \dots X_p | Y = i)$, into parts that can be expressed as follows: $P(X_1 | Y = 0) \dots P(X_p | Y = 0)$. The formula for Naive Bayes is then as follows:

$$P(Y = i | X_1, X_2, \dots, X_p) = \frac{P(Y=i) \cdot P(X_1|Y=i) \cdot \dots \cdot P(X_p|Y=i)}{P(Y=0) \cdot P(X_1|Y=0) \cdot \dots \cdot P(X_p|Y=0) + P(Y=1) \cdot P(X_1|Y=1) \cdot \dots \cdot P(X_p|Y=1)}$$

The assumption that these parameters are independent is, however, unrealistic and for this reason, the algorithm is referred to as Naive Bayes. It should be noted that the formulas described thus far are only applicable to categorical data. In order for the algorithm to be effective when working with numeric attributes, it is necessary to either bin and convert the data into a categorical format or to assume that the attributes follow a normal distribution or some other distribution.

2.6. Neural Networks

In the past decades, the artificial neural network (neural networks for short) algorithm has taken the world by the storm. Besides being able to achieve state of the art performance on classical classification and regression tasks, neural networks also serve as the foundation for convolution and recurrent neural networks which are achieving remarkable results in image and text

processing, respectively. Furthermore, neural networks lie in the heart of transformers, models that have enabled a recent revolution in Generative AI, as evidenced by ChatGPT, LLaMA, and Grok, among other examples. The name of the algorithm is derived from the biological nervous system, as its functioning has been inspired by this model.

Prior to an examination of artificial neurons, it is essential to establish a clear understanding of the concept of a biological neuron. A nerve cell represents the fundamental unit of structure and function within the nervous system. A neuronal cell is comprised of cytoplasm, which contains a nucleus and a multitude of complex components. Additionally, it comprises dendrites, which are short nerve endings on the neuron's body, and an axon, which is a long extension of the cytoplasm. A biological neuron receives electrical impulses from other neurons, which are transmitted via synapses. Synapses are structures situated between the dendrites of one neuron and the axonal terminal of another, facilitating communication between neurons. When a neuron receives a sufficient number of signals from other neurons, it becomes activated and transmits its signal to other neurons [23].

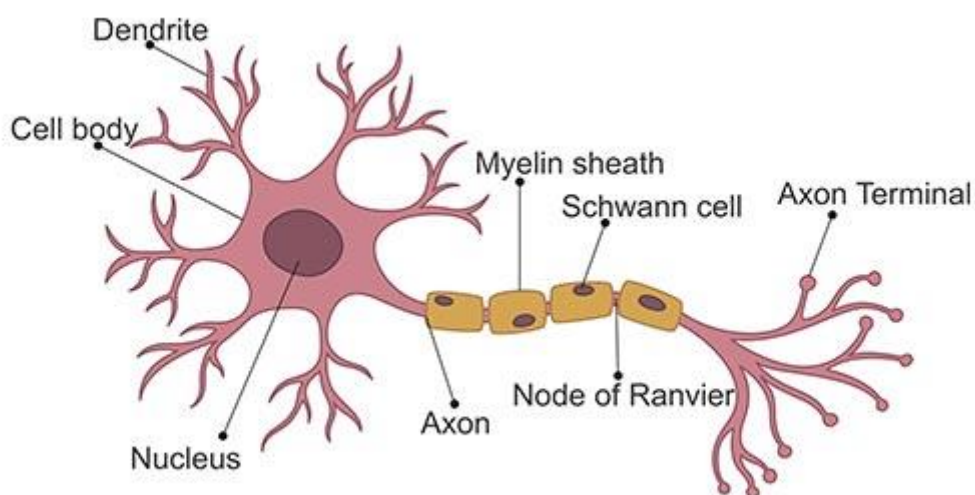


Figure 2.9 The Typical Structure of a Neuron [27]

Biological neurons operate in a relatively straightforward manner, yet they are organised into extensive networks comprising billions of neurons, with each neuron linked to thousands of others. The architectural configuration of biological neural networks remains incompletely elucidated; nevertheless, the components that have been investigated indicate that neurons are arranged in successive layers, as illustrated in the figure below. This architectural model has prompted researchers to develop artificial neural networks.

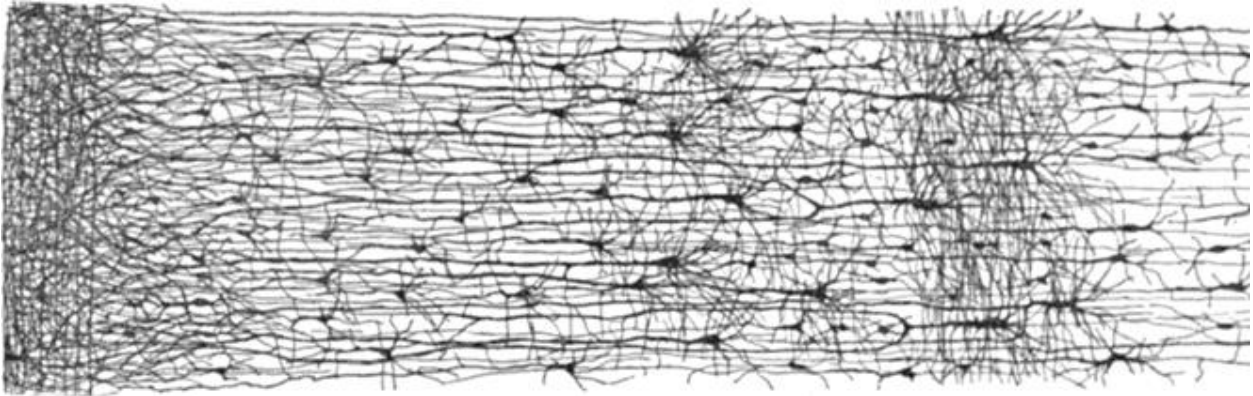


Figure 2.10 A representation of the multiple layers that constitute a biological neural network, as illustrated by S. Ramon y Cajal in his study of cortical lamination [23]

Artificial neural networks are composed of multiple layers. Each network begins with an input layer, which serves to introduce initial data into the network. This layer represents the initial stage of the network's operational process. The output layer represents the result calculated by the neural network. The hidden layers, situated between the input and output layers are where the training and computational processes occur.

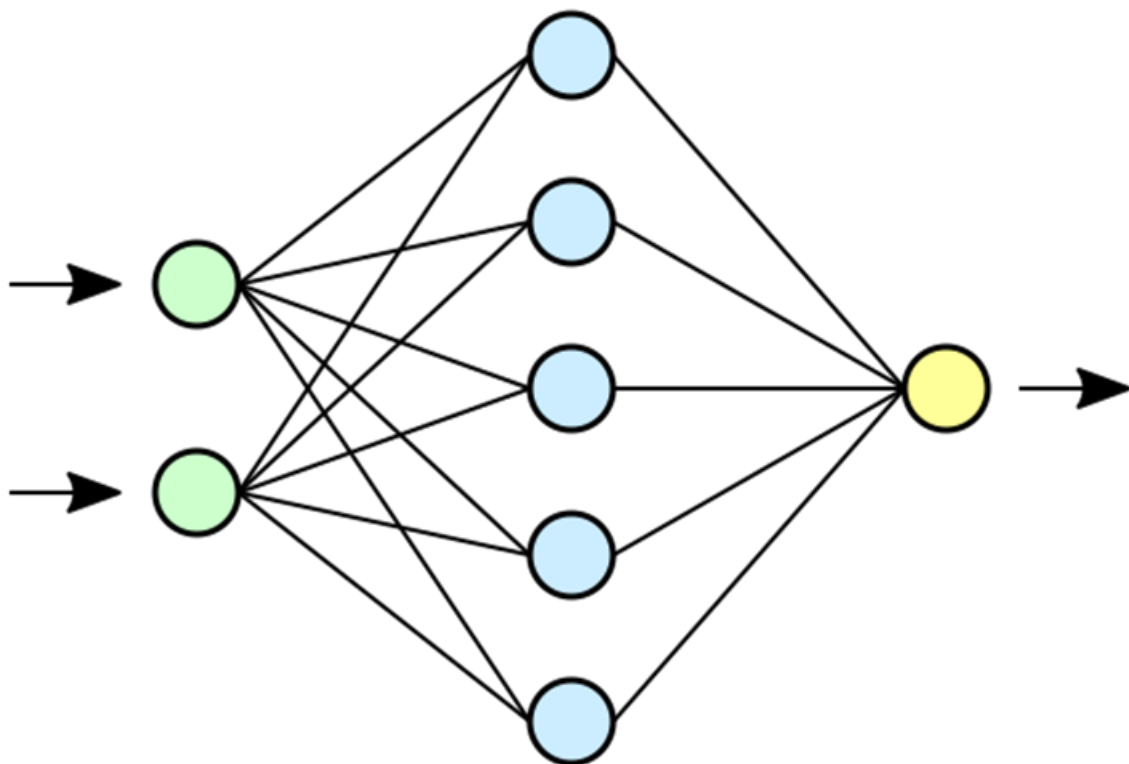


Figure 2.11 The architecture of the neural network is presented as follows: the input layer is depicted in green, the hidden layer in blue, and the output layer in yellow [23]

A neural network is composed of a series of interconnected layers, each of which consists of multiple nodes, or neurons. A neuron in a neural network strives to emulate its biological counterpart through the utilisation of mathematical functions, but unlike the biological neuron,

which exhibits binary activity, the activity of an artificial neuron is represented by a decimal number usually situated between zero and one. Each neuron is connected to all the others in the previous layer through connections, known as weights, which indicate the extent to which the activation of one neuron depends on the activation of another. In technical terms, weights quantify the connection between two neurons. To calculate the activation of each, the matrix product of the weights and the corresponding neurons is computed, a bias term is added, and then a nonlinear function is applied.

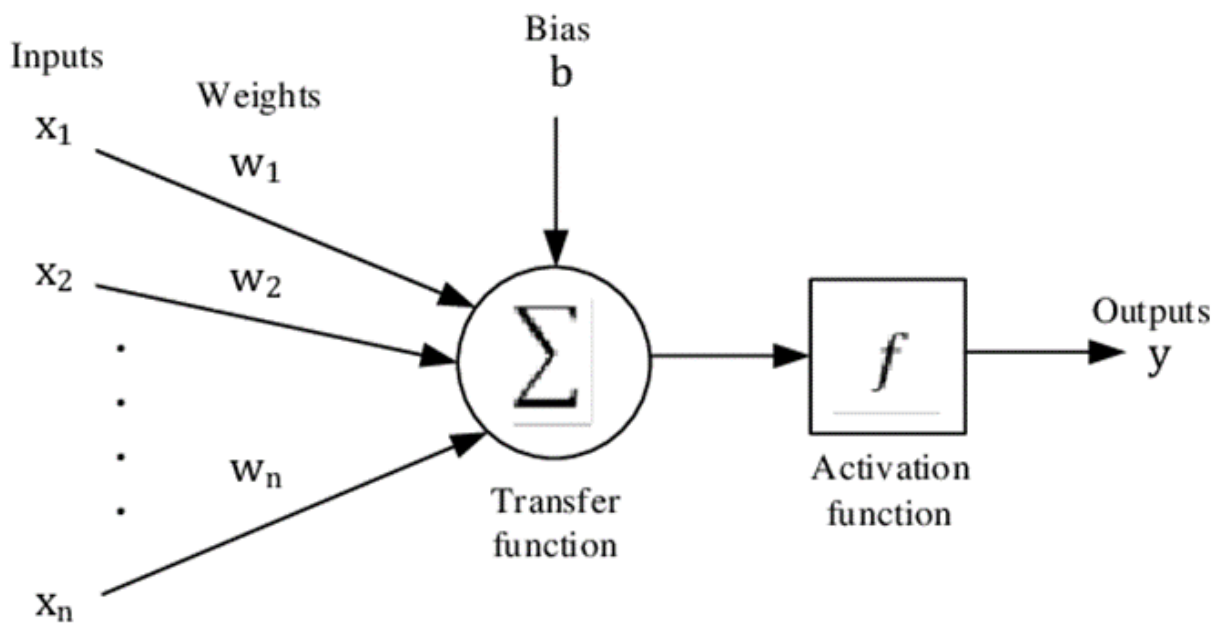


Figure 2.12 Image showing a computation of a value for a single neuron. [23]

The addition of a nonlinear function introduces complexity to the model, which is advantageous for classification and regression problems. This is because the solution to the majority of real-world problems cannot be expressed as simple addition and multiplication of numbers. Furthermore, the training of a neural network with more than one layer is not possible without a nonlinear function. This is because successive linear operations can be expressed as a single one.

A nonlinear function that is frequently employed in neural networks is the sigmoid function. However, researchers have found that the ReLU activation function (Rectified Linear Unit) and its variations yield superior results. The ReLU activation function is differentiable and monotonic, and, in contrast to the majority of other activation functions, its derivative is also monotonic. The function converts all negative values to zero, thereby indicating that all insignificant values are equally insignificant. ReLU is currently the most widely used activation function, particularly in deep and convolutional neural networks.

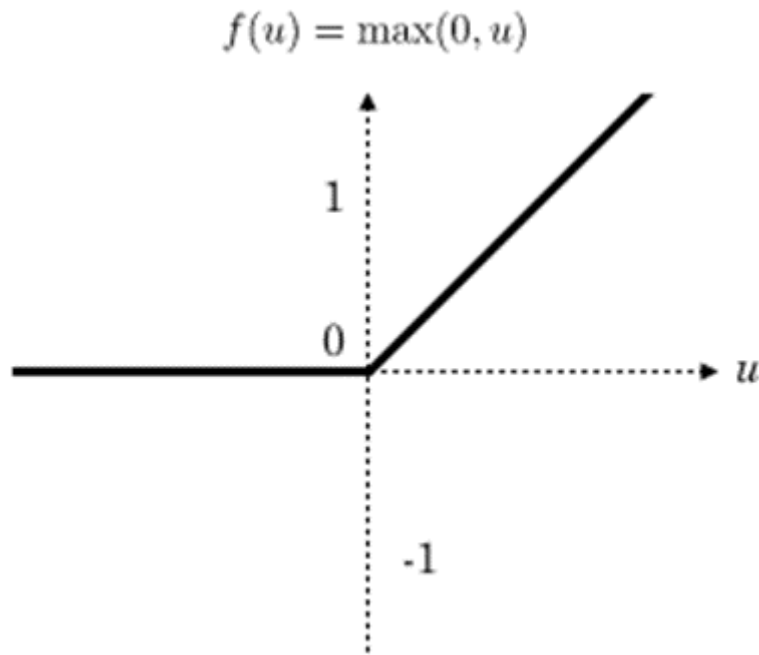


Figure 2.13 ReLU activation function [23]

Due to its property of outputting values between 0 and 1, the sigmoid function is used in the binary classification output layers to produce the probability of belonging to a specific class. Its formula is expressed as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

The process by which a neural network computes the output based on the input data is referred to as forward propagation. This involves a series of matrix multiplications, followed by the addition of nonlinearity. The activation values in the initial layer are equal to the input data. They are then multiplied by the weights, and a bias is added. Subsequently, an activation function is applied, resulting in the activation values for the second layer. This process is repeated until the final layer, and it follows the following formulas:

$$Z^{(2)} = W^{(1)} \cdot X + b^{(1)}$$

$$a^{(2)} = \sigma(Z^{(2)})$$

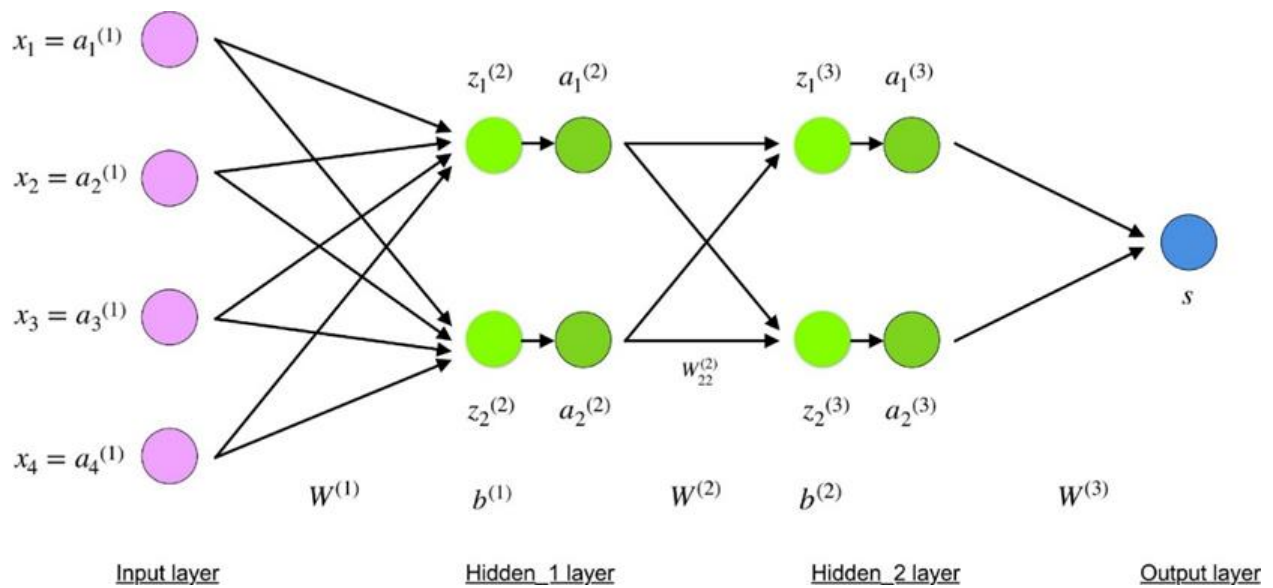


Figure 2.14 The architecture of a neural network with activations [23].

Upon reaching the final layer of the process, the algorithm produces predictions for a given feature vector. These are then compared with the true value, and a loss function is computed. The specific loss function employed depends on the task at hand. The binary cross-entropy loss function, which is commonly used in binary classification tasks, measures the difference between two probability distributions: the true distribution (labels) and the predicted distribution (output from a model). Given the true labels, represented by y , which take on the values 0 or 1, and the predicted probabilities, represented by the vector of probabilities, denoted by the symbol \hat{y} , with elements in the interval $[0, 1]$, the binary cross-entropy loss is:

$$C(y, \hat{y}) = -(y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y}))$$

The model is trained based on the loss function and the process known as backpropagation. Backpropagation is an iterative process that adjusts the weights of a neural network until the model reaches a global minimum or meets another pre-defined stopping criterion. The method calculates the contribution of each parameter to the error and adjusts the weights accordingly. The term 'backpropagation' is derived from the fact that the weight updates begin at the last hidden layer and move backward to the first hidden layer. The magnitude of the adjustment for a given weight is dependent on the weights present between it and the output.

The chain rule forms the foundation for backpropagation. In the field of calculus, the chain rule is a mathematical formula that expresses the derivative of the composition of two differentiable functions, designated as f and g , in terms of the derivatives of f and g . When we have $y = f(g)$ and $u = g(x)$, the derivative of y with respect to x is:

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

The same process is applied to neural networks, which can also be viewed as a composite function. The subsequent image demonstrates the calculation of derivatives, which, when calculated in higher dimensional space, are referred to as gradients.

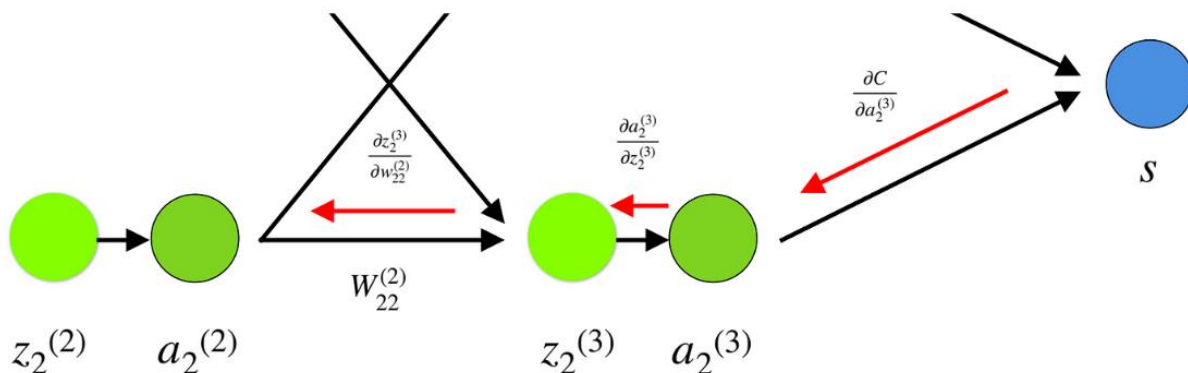


Figure 2.15 A part of the neural network illustrating partial derivatives and backpropagation [23].

Gradients are calculated because they show the direction of the biggest ascent/descent. Once the path with the greatest descent has been identified, the weights can be adjusted in accordance with the formula presented below:

$$w := w - \alpha \nabla_w C(w)$$

It is evident that prior to undertaking any adjustments to the weights with a view to enhancing the accuracy of the model, it is first necessary to have some initial weights to modify. The process of assigning weights prior to the onset of training is referred to as weight initialization. It is of paramount importance that this process is conducted correctly, as any errors may result in the algorithm failing to learn effectively. If all weights are initialised to zero, if the initialization is symmetric, or if there is inconsistent variance between layers, the training will not be optimal. The most effective method for initialization is the Xavier initialization, whereby the values for each weight are drawn from a uniform distribution bounded by the following values:

$$\text{Uniform} \left(-\sqrt{\frac{6}{n_{\text{in}} + n_{\text{out}}}}, \sqrt{\frac{6}{n_{\text{in}} + n_{\text{out}}}} \right)$$

where n_{in} is the number of input neurons to the layer, while n_{out} denotes the number of output neurons from the layer. This initialization process serves to maintain the variance of the activations and gradients across layers, which in turn leads to more stable and efficient training.

2.7. K-Nearest Neighbours

The K-nearest neighbours algorithm is a supervised learning algorithm that can be utilised for both regression and classification purposes. It is an instance-based model, meaning that it learns training data by heart and, when predicting a new outcome, it compares the input point to the closest K points in the metric space. Instance-based models differ from model-based ones in that the latter have several predetermined parameters, whereas the former do not. In the case of

linear regression, for example, which is a model-based algorithm, the slope of the line and the bias are learned, but the data on which it is trained are not stored.

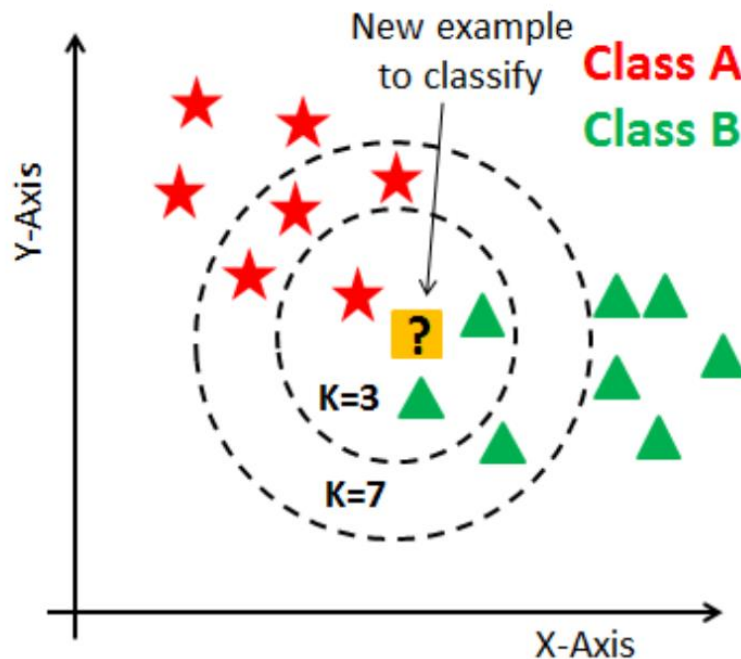


Figure 2.16 The image shows how classifying new examples works. It also shows that depending on parameter K different predictions are obtained [29].

In the event of regression, the algorithm calculates the mean value of the target values of the nearest points. In the case of classification, the algorithm identifies the most prevalent class among its neighbouring data points. In both cases, the algorithm has the option of weighting all neighbours equally or weighting them based on distance. The choice of space in which distances are measured also has a significant impact. Two of the most commonly employed distance metrics are the Euclidean distance or L2 norm and the Manhattan distance or L1 norm. The Euclidean distance assigns greater weight to instances that are distant along a single axis, in contrast to the Manhattan distance, which considers only the absolute distance along each axis. The most crucial hyperparameter is the number of neighbours. As illustrated in the above image, the classifier's behaviour varies depending on the number of neighbours. Therefore, it is essential to evaluate different values of K and identify the optimal one for the validation data.

K-NN is frequently employed as a preprocessing step to estimate missing values, as it is capable of identifying the nearest rows that do not contain the value in question. It may be employed in recommendation engines and loan approval in situations where there are no extensive datasets. For instance, one paper [22] illustrates how the use of KNN on credit data can assist banks in assessing the risk of a loan to an organisation or individual. Additionally, KNN has been utilised in the healthcare industry, enabling the prediction of the risk of heart attacks and prostate cancer. The algorithm operates by calculating the most probable gene expressions [21].

The principal advantage of the K-Nearest Neighbours algorithm is that it is straightforward to implement and comprehend. In comparison to other machine learning algorithms, K-Nearest Neighbours has only two hyperparameters, and it is easily implemented in an online mode,

allowing for the straightforward addition of new data to the model without the need for retraining. The primary disadvantages of K-NN are its requirement for the storage of all data, which is impractical for larger datasets, and its susceptibility to the curse of dimensionality, whereby its distance metric loses value in higher dimensions. Additionally, K-NN is known to present challenges in the context of imbalanced datasets.

3. Data Science in Medicine

Healthcare has a long history of using data analysis techniques to understand disease and develop cures. And to understand the current state of affairs, one needs to look at the history. In this chapter, we will briefly review some of the most notable moments in the development of data science in medicine. These stories help to explain the foundations of the field and where advanced methods have come from. At the end, we will explain how things stand today and how they were applied during the COVID-19 pandemic.

One of the earliest instances of a controlled clinical trial (an example of A/B testing) was conducted by Scottish physician James Lind in 1747. He was aboard a British vessel that was carrying numerous sailors afflicted with scurvy. Lind proposed the hypothesis that the lack of citrus fruit may be the reason for the occurrence of scurvy, given that sailors on Mediterranean ships did not experience this condition. Lind therefore provided limes to half of his sailors, while the remaining half continued with their normal diet. In statistical terms, these groups are referred to as the 'treatment' and 'control' groups, respectively. Lind's hypothesis was confirmed, as the sailors who consumed limes exhibited improvement. However, the captain was unaware that scurvy is a consequence of vitamin C deficiency and that limes are a rich source of this vitamin. Nevertheless, British sailors were eventually compelled to consume citrus fruit on a regular basis.

Another individual who contributed to the popularisation of the use of statistics in the field of medicine was Florence Nightingale. She was an attentive nurse during the Crimean War, during which she had the opportunity to observe the horrific effects of war first-hand. The most disturbing aspect of the situation was the deplorable living conditions of the soldiers, who were deprived of basic necessities such as fresh air and clean water. Nightingale observed that a greater number of deaths were occurring from preventable diseases than from deadly wounds, and she collected data in order to prove this hypothesis. Rather than sharing her observations in the conventional table format, Nightingale drew visualisations of the data in order to facilitate comprehension by a wider audience, as she was engaged in the campaign to promote sanitary reform. The efforts she had invested were successful, and parliament passed the British Public Health Act of 1875, which established requirements for well-built sewers and clean running water.

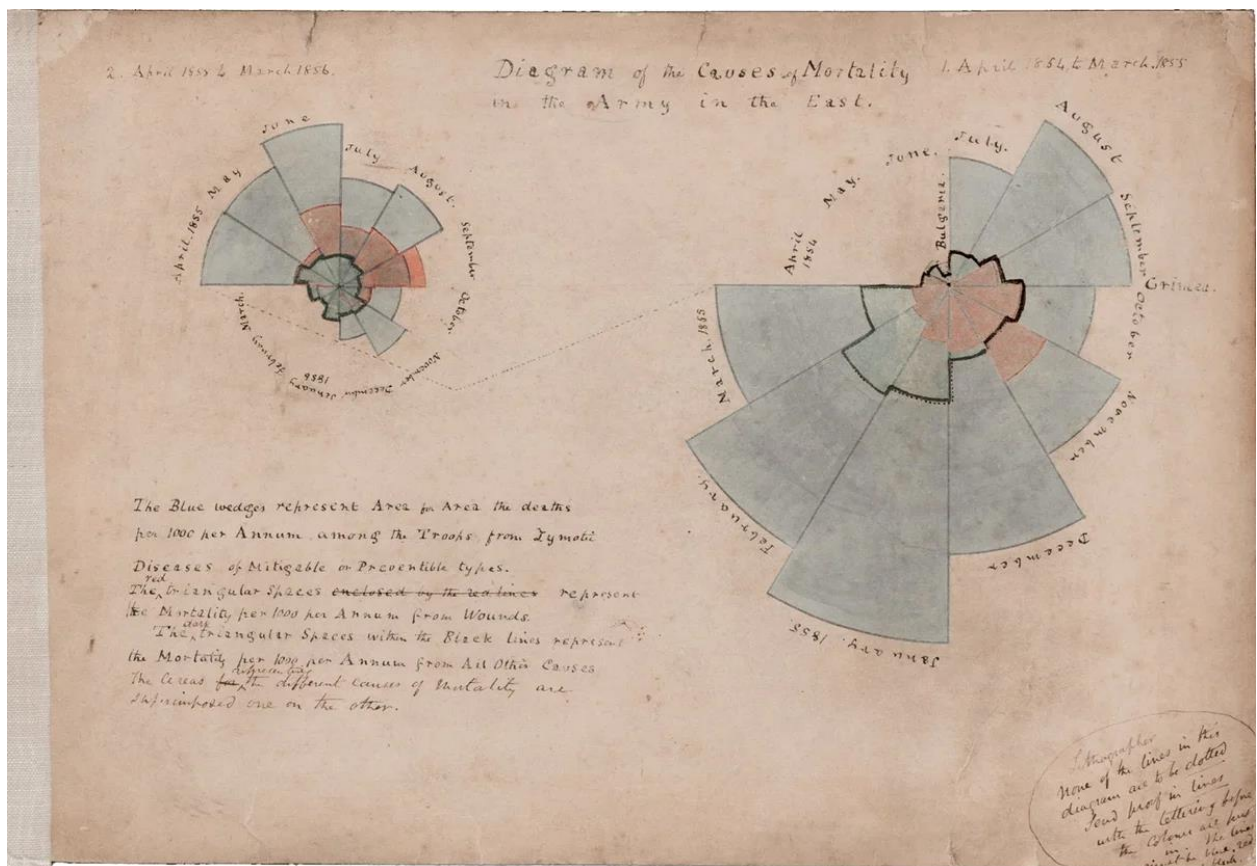


Figure 3.1 The surviving drafts of Nightingale's diagrams illustrate a discrepancy between the number of army deaths from preventable diseases (blue) and the number of hospital deaths from wounds (red) [7].

Nevertheless, statistical methods were not always readily accepted by the medical profession. One such case was the proposal by Ignaz Semmelweis that doctors should wash their hands before assisting with the delivery of a baby. Although this may appear to be self-evident to the modern reader, it was not so to the doctors of the time, who did not wash their hands. Semmelweis observed that poor hygiene was a significant factor in the mortality of women and made this observation by comparing the death rates of two hospitals: one with midwives exclusively attending to women in labour, and the other with medical students who also spent time in the autopsy rooms examining corpses. It is perhaps unsurprising that the hospital with the higher death rate was the one with medical students in attendance.

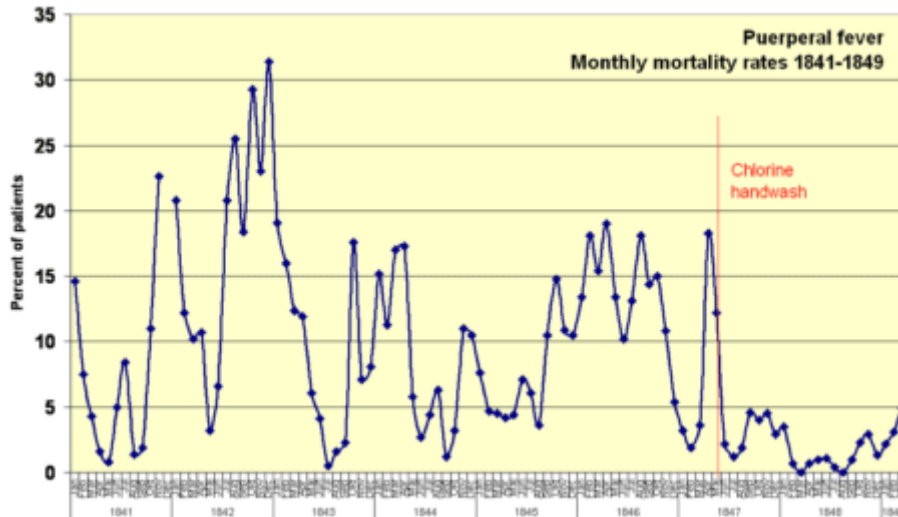


Figure 3.2 Monthly mortality rates 1841-1849 in the hospital Semmelweis led. That was gathered by the doctor but he did not put it into a visual form [28].

Semmelweis collated the relevant data and demonstrated that this was indeed the case. He introduced the mandatory practice of hand washing and subsequently documented the number of deaths that occurred with this procedure in place. The mortality rate fell from an average of 10% to only 2%. Despite the evidence, doctors rejected Semmelweis' findings and even ridiculed him. They were reluctant to accept that they were responsible for the excess deaths. In 1849, he was forced to leave the Vienna hospital. He subsequently experienced a nervous breakdown and was admitted to a mental hospital by his peers. He was subjected to physical abuse and died shortly afterwards.

Today, the application of AI is being explored across a range of levels, from molecular to population-level. The healthcare sector is at the forefront of utilising data science to drive innovation, enhance operational effectiveness, provide precision medicine solutions and optimise patient outcomes. The healthcare sector has a long tradition of utilising research and in-depth data analysis to generate new insights into disease progression, drug development and other areas. The application of advanced data science technologies has the potential to accelerate these processes, enhance their precision, and extend their scope to larger scales. The advent of Big Data and machine learning algorithms has enabled the rapid analysis of extensive DNA sequences. Models have been developed to predict the probability of an elderly patient with migraine experiencing a stroke, and convolutional neural networks have been employed to facilitate the inspection of radiology scans.

In the aftermath of the global pandemic caused by the SARS-CoV-2 virus, the field of data science has experienced a surge in popularity. The significance of this technology is evident in its impact on a diverse range of individuals and groups, including researchers, healthcare professionals, policymakers, academics, decision-makers, and the general public. The technology was employed for a number of purposes, including:

- The creation of accessible yet informative data visualisations and dashboards
- The identification of the subsequent surge in cases of coronavirus

- The prediction of immunity to, and risk of infection from, the virus
- The acceleration of the discovery of treatments
- The assessment of the economic impact and facilitation of changes

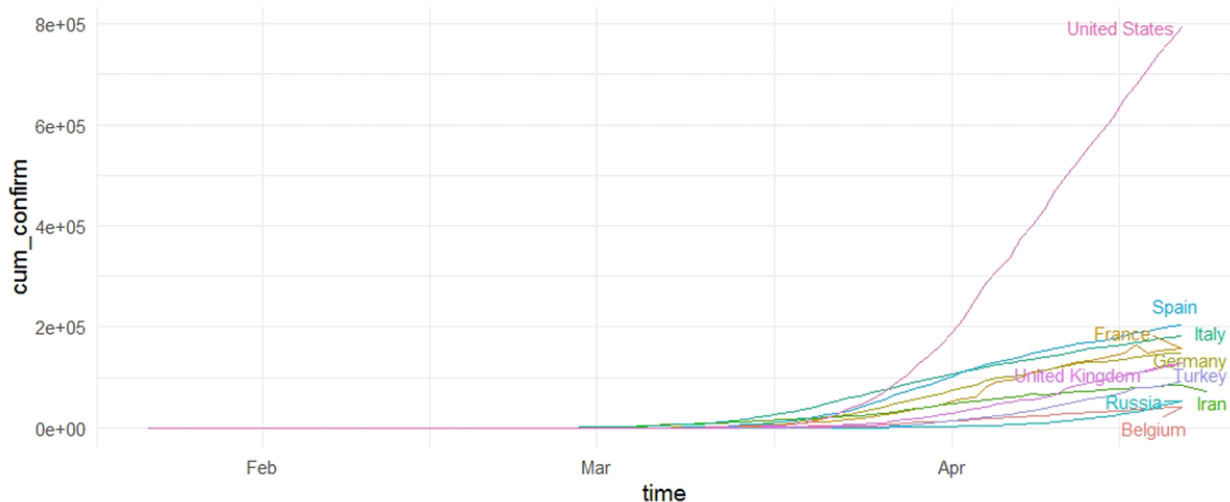


Figure 3.3 One illustrative example of a significant data visualisation employed during the global pandemic is the presentation of the total number of cases worldwide, with the exception of China.

The healthcare sector has witnessed rapid growth in the utilisation of AI technology. Nevertheless, there remains untapped potential for AI to truly revolutionise the industry. It is important to note that despite concerns about job displacement, AI in healthcare should not be viewed as a threat to human workers. Instead, AI systems are designed to augment and support healthcare professionals, freeing up their time to focus on more complex and critical tasks. By automating routine and repetitive tasks, AI can alleviate the burden on healthcare professionals, allowing them to dedicate more attention to patient care and meaningful interactions. However, legal and ethical challenges must be addressed when embracing AI technology in medicine, alongside comprehensive public education to ensure widespread acceptance [26].

4. Related work

This master's thesis is largely based on the work of Mitrovic, Pantic, Bukumirc et al., entitled "Venous thromboembolism in patients with acute myeloid leukaemia: development of a predictive model". A retrospective cohort study was conducted on adult patients with newly diagnosed acute myeloid leukaemia. The data were gathered at the Clinic for Hematology at the University Clinical Center of Serbia between 2009 and 2021. The researchers employed univariate and multivariable logistic regression to estimate binary outcomes and identify potential predictors. They identified five predictors that were statistically significant: patient sex, prior history of thrombotic events, international normalised ratio (INR), Eastern Cooperative Oncology Group performance status (ECOG) and intensive therapy. The area under the curve (AUC) statistics of the final model was 0.68.

As this paper was designed to identify the causes of thrombosis and was intended for a medical audience, the authors chose to avoid incorporating modern data science concepts. Firstly, parameters that were on vastly different scales were not subjected to preprocessing in order to ensure that they were within the same range. Secondly, the data was not divided into a training set and a test set; rather, the model was trained on the entire dataset, and the results were reported on the entire dataset. This may not be problematic when working with logistic regression, which is linear in nature. However, when using more complex non-linear models, it is advisable to use a test set as a guard against overfitting. Thirdly, an extensive parameter search was not performed, meaning that regularisation and related parameters were not tried.

5. Research methodology

The objective of this thesis is to investigate the potential of statistical and machine learning models for predicting venous thrombosis in patients with acute myeloid leukaemia. Given that thrombosis is a significant contributor to morbidity and mortality among patients with cancer, particularly those with leukaemia, effective treatment strategies are of paramount importance. However, the use of thromboprophylaxis, or preventive measures to reduce the risk of thrombosis, is limited by a high prevalence of thrombocytopenia and the perceived high risk of bleeding, as well as the lack of evidence-based guidelines to assist clinicians. As a result, determining the risk factors for VTE in patients with acute leukemias will enable clinicians to risk-stratify patients and individualise patient surveillance and preventive blood-thinning treatment.

The data collection process included the following variables: demographic factors (age, sex), body mass index (BMI), smoking status, comorbidities (including previous thrombosis), concomitant therapy, ECOG PS, Hematopoietic Cell Transplantation-specific Comorbidity Index, and baseline laboratory findings (complete blood count, fibrinogen, prothrombin time [PT], International Normalized Ratio [INR], activated partial thromboplastin time [APTT]). Additionally, the following parameters were assessed: APTT, D-dimer, lactate dehydrogenase (LDH), leukemia-related parameters (cytogenetics, molecular genetics [FLT3, NPM1], flow cytometry), type (intensive, non-intensive, palliative therapy) and phase of leukemia-related therapy, the presence of a central venous line (CVL), Khorana and AI Ani scores, and concurrent coronavirus disease 2019 (COVID-19) positivity. DIC was diagnosed in accordance with the International Society on Thrombosis and Haemostasis (ISTH) scoring system. All laboratory parameters, as well as comorbidities, concomitant therapy, and smoking status, were assessed on the day of diagnosis or the nearest day before, within a three-day period.

Parameter	All (n=626)	Missing values (%)	Patients with thrombosis (n=72)	Patients without thrombosis (n=554)	OR	95% CI	p-value
Age (years)	55.1±13.4	0	52.9±13.7	55.4±13.3	0.99	0.97–1.004	0.137
Male sex (%)	348 (55.6)	0	49 (68.1)	299 (54.0)	1.82	1.08–2.07	0.025
Smokers (%)	277 (46.8)	16	35 (51.5)	242 (46.2)	1.24	0.75–2.05	0.412
BMI	25.3±4.7	7.7	25.6±4.0	25.2±4.8	1.01	0.96–1.07	0.598
Prior history of thrombotic events (%)	42 (6.8)	12.5	9 (12.7)	33 (6.9)	2.27	1.04–4.96	0.041

ECOG PS2 (%) 0	102 (16.7)	13.3	17 (25.0)	85 (15.7)	0.71	0.53–0.94	0.017
ECOG PS2 (%) 1	256 (41.9)		30 (44.1)	226 (41.6)			
ECOG PS2 (%) 2	182 (29.8)		17 (25.0)	165 (30.4)			
ECOG PS2 (%) 3	48 (7.9)		3 (4.4)	45 (8.3)			
ECOG PS2 (%) 4	23 (3.8)		1 (1.5)	22 (4.1)			
Comorbidities Total number	1 (0–7)	1.9	1 (0–4)	1 (0–7)	0.85	0.67–1.08	0.193
Diabetes (%)	102 (17.4)	17.4	10 (14.5)	92 (17.8)	0.78	0.39–1.59	0.498
Hypertension (%)	156 (25.0)	17.4	14 (20.3)	142 (27.5)	0.67	0.36–1.25	0.208
Antiplatelet therapy (%)	33 (5.4)	1.9	5 (7.1)	28 (5.1)	0.71	0.26–1.89	0.488
HCT Cl3 (%)	1 (0–9)	2.4	1 (0–4)	1 (0–9)	0.83	0.83?0.69	0.052
Khorana score (%) 0	112 (17.9)	11.5	13 (18.1)	99 (17.9)	0.94	0.66–1.32	0.708
Khorana score (%) 1	322 (51.4)		39 (54.2)	283 (51.1)			
Khorana score (%) 2	184 (29.4)		19 (26.4)	165 (29.8)			
Khorana score (%) 3	8 (1.3)		1 (1.4)	7 (1.3)			
AI Ani score (%) 0	317 (50.6)	0	30 (41.7)	287 (51.8)	1.26	0.90–1.78	0.185
AI Ani score (%) 1	296 (47.3)		40 (55.6)	256 (46.2)			
AI Ani score (%) 2	0 (0.0)		0 (0.0)	0 (0.0)			

AI Ani score (%) 3	8 (1.3)		2 (2.8)	6 (1.1)			
AI Ani score (%) 4	5 (0.8)		0 (0.0)	5 (0.9)			
COVID-19 (%)	59 (9.4)	11.6	7 (9.7)	52 (9.4)	1.04	0.45–2.38	0.931
CNS involvement (%)	54 (20.5)		11 (30.6)	43 (18.9)	1.89	0.87–4.14	0.11
WBC (normal: 3.6–10×10 ⁹ /L)	9.8 (0.4–473.2)	0	10.5 (0.7–211.6)	9.7 (0.4–473.2)	0.998	0.993–1.002	0.321
Platelet count (normal: 150–400×10 ⁹ /L)	49 (1-726)	0	56 (1-220)	47 (1-726)	1.001	0.998–1.004	0.37
Hemoglobin (normal: 120–160 g/L)	95.8±17.8	0	97.0±18.8	95.7±17.4	1.004	0.991–1.018	0.542
LDH (normal, 220–460 U/L)	458 (105–8902)	9.4	384 (180–4150)	465 (105–8902)	1	0.999-1.000	0.17
Fibrinogen (normal: 2.2–5.5 g/L)	5.4 (0.3–56.0)	5.2	5.6 (1.4–8.5)	5.3 (0.3–56.0)	0.928	0.821–1.048	0.229
INR (normal: 0.8–1.3%)	1.22±0.19	5.2	1.18±0.17	1.23±0.20	0.21	0.05–0.95	0.043
APTT (normal: 25.1–36.5 s)	29.2±5.6	5.2	28.4±4.2	29.3±5.7	0.96	0.91–1.02	0.198
D dimer (normal: 0–0.5 µg/L)	2.5 (0.1–158.0)	26.5	2.1 (0.3–100.8)	2.5 (0.1–158.0)	0.99	0.98–1.01	0.649

ISTH DIC score ²⁵ (%)	131 (41.3)	26.5	12 (28.6)	119 (43.3)	0.52	0.26–1.07	0.075
Blast peripheral blood (%)	16 (0–99)		15 (0–98)	17 (0–99)	0.99	0.98–1.003	0.182
FAB (%) 0	32 (5.3)	3.4	3 (4.3)	29 (5.4)	0.99	0.90–1.10	0.881
FAB (%) 1	69 (11.4)		12 (17.4)	57 (10.6)			
FAB (%) 2	150 (24.8)		19 (27.5)	131 (24.4)			
FAB (%) 3	2 (0.3)		0 (0.0)	2 (0.4)			
FAB (%) 4	172 (28.4)		16 (23.2)	156 (29.1)			
FAB (%) 5	99 (16.4)		7 (10.1)	92 (17.2)			
FAB (%) 6	2 (0.3)		0 (0.0)	2 (0.4)			
FAB (%) 7	1 (0.2)		0 (0.0)	1 (0.2)			
FAB (%) 9	78 (12.9)		12 (17.4)	66 (12.3)			
ELN classification (%) Good	66 (11.4)		8 (11.9)	55 (11.3)	0.88	0.58–1.33	0.529
ELN classification (%) Intermediate	330 (59.5)		42 (62.7)	288 (59.0)			
ELN classification (%) High	162 (29.2)		17 (25.4)	145 (29.7)			
FLT3 ITD positivity (%)	63 (19.9)	-	9 (20.9)	54 (19.7)	1.08	0.49–2.38	0.852
NPM1 positivity (%)	59 (24.4)	-	11 (33.3)	48 (23.0)	1.68	0.76–3.70	0.201
CD56 positivity (%)	175 (33.1)	22.3	19 (29.7)	156 (33.5)	0.84	0.47–1.48	0.539

CD13 positivity (%)	510 (93.1)	22.3	59 (90.8)	451 (93.4)	0.7	0.28–1.74	0.44
CD34 positivity (%)	382 (69.5)	22.3	42 (64.6)	340 (70.1)	0.78	0.45–1.34	0.368
CD33 positivity (%)	512 (93.1)	22.3	60 (90.9)	452 (93.4)	0.71	0.28–1.76	0.458
CD117 positivity (%)	482 (87.8)	22.3	55 (87.8)	427 (87.9)	0.95	0.43–2.09	0.899
CD7 positivity (%)	126 (23.8)	22.3	11 (17.5)	115 (24.6)	0.65	0.33–1.28	0.213
CD15 positivity (%)	178 (34.0)	22.3	21 (33.9)	157 (34.0)	0.99	0.57–1.74	0.986
CD19 positivity (%)	49 (9.5)	22.3	6 (9.8)	43 (9.5)	1.04	0.42–2.56	0.932
CVL inserted (%)	519 (82.9)	11.5	68 (94.4)	451 (81.4)	3.88	1.38–10.89	0.01
Therapy type - Intensive (%)	453 (72.4)	11.3	60 (83.3)	393 (70.9)	2.05	1.07–3.91	0.03
Therapy type - Non-intensive (%)	173 (27.6)	-	12 (16.7)	161 (29.1)			

Table 5.1 The summarisation of the entire dataset

A review of the collected data revealed a significant number of columns with missing information. In practice, it is not uncommon for some cases to be excluded from the data set due to the unavailability of information for various reasons. The conventional approach to dealing with missing data is to exclude any column in which more than 5% of the data is missing. The columns designated as 'CNS.inf.likvor', 'fit3_ITD', 'NPM1', 'CD56', 'CD13', 'CD34', 'CD33', 'CD117', 'CD7', and 'CD15' The variables 'CD19', 'D.dimer', 'Comorbidities', 'ISTH DIC score', 'COVID19', and

'Khorana.score' were excluded from further analysis due to the presence of an excess of null values.

Prior to undertaking any additional preprocessing, it is essential to divide the data into training and test sets, as the latter must remain intact until the model has been validated. Otherwise, information about it would be leaked, resulting in biased validation. One example of data leakage is mean value imputation, whereby the mean is calculated on the entire dataset. Subsequently, when the data is partitioned, one could infer the mean values of features in the test set by calculating the mean value of features in the training set. The data split was 70/30, and the data was stratified so that there were equal proportions of positive cases in both the training and testing datasets.

Afterwards, all features were scaled to a uniform range using the scikit-learn MinMaxScaler function. Scaling specifications were determined on the training dataset and subsequently applied to both the training and test sets. Further feature reduction was then conducted by removing those features that were dependent on other features. To assess this, Pearson's correlation coefficient and variance inflation factor were employed. The formula for Pearson's coefficient is:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

- X_i and Y_i are individual data points for variables X and Y .
- \bar{X} and \bar{Y} are the mean values of X and Y , respectively.
- r ranges from -1 to 1:
- $r = 1$ indicates a perfect positive linear correlation.
- $r = -1$ indicates a perfect negative linear correlation.
- $r = 0$ indicates no linear correlation

The subsequent image illustrates the correlation between the remaining columns. It can be observed that there is a strong positive correlation between "Tip.terapije" (type of therapy) and CVK; therefore, CVK was excluded.

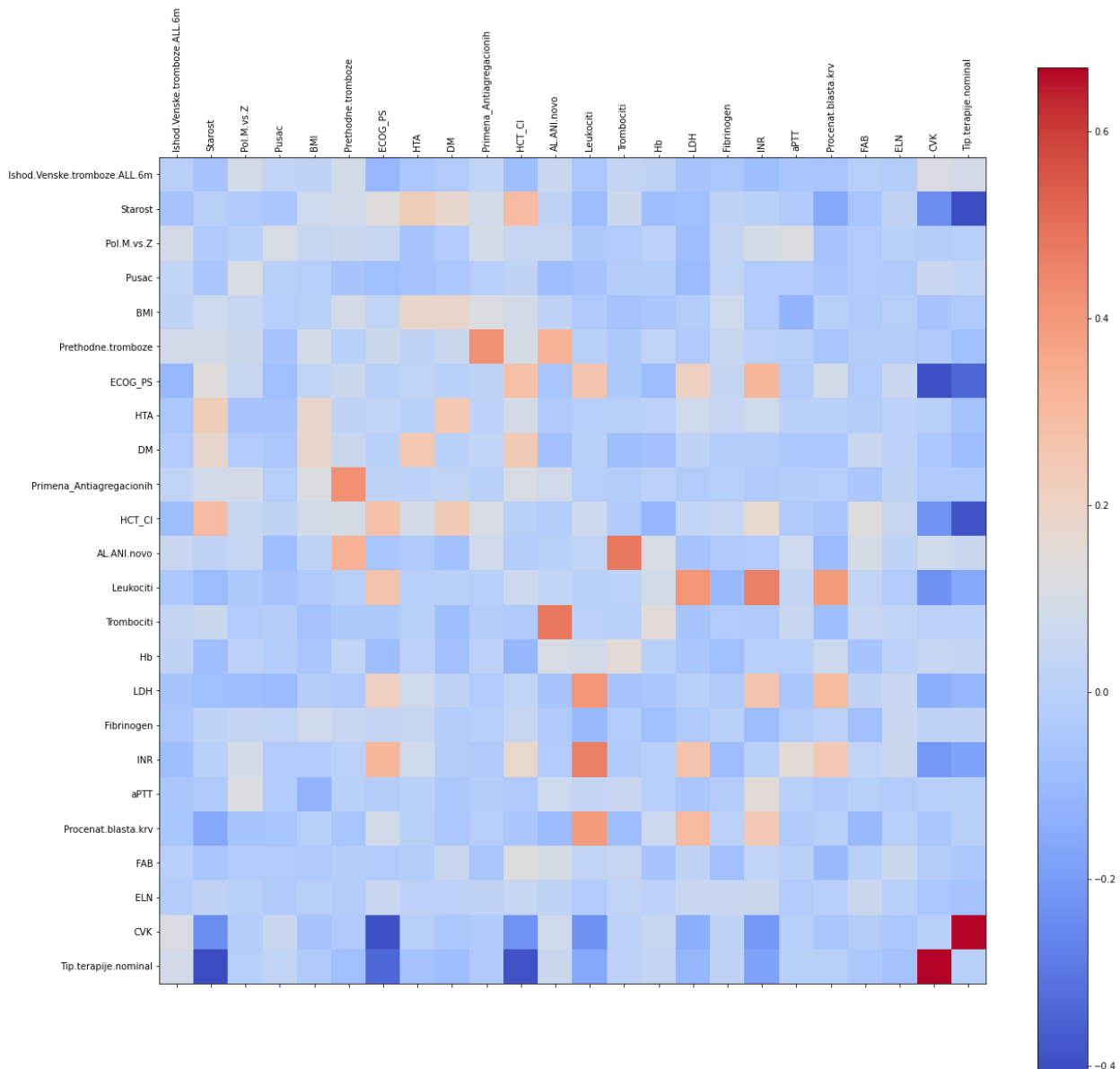


Figure 5.1 The image depicts a matrix of linear correlations, with each cell coloured according to the strength of the correlation between the corresponding row and column attributes.

The Variance Inflation Factor (VIF) is a statistical measure employed in regression analysis to identify instances of multicollinearity among predictor variables. Multicollinearity arises when two or more predictors in a model exhibit a high degree of correlation, which can result in the generation of unreliable estimates of regression coefficients. The Variance Inflation Factor (VIF) is a statistical measure used to quantify the extent to which the variance of a regression coefficient is inflated due to collinearity with other predictor variables. It is calculated for each predictor variable by regressing it against all other predictor variables and determining the extent to which the variance of the coefficient is increased. A high VIF value indicates a high level of multicollinearity, which is typically defined as a VIF value greater than 10, suggesting problematic multicollinearity that may require further investigation or adjustment to the model.

The following is a brief overview of the Variance Inflation Factor (VIF) values for each variable in

the dataset.

Variable	VIF
Age	1.413224
Sex	1.076852
Smokers	1.048541
BMI	1.108553
Prior history of thrombotic events	1.476636
ECOG_PS	1.324729
HTA	1.160239
DM	1.183053
Antiplatelet therapy	1.266924
HCT_CI	1.381181
AL.ANI. new	1.613094
Leukocytes	1.647233
thrombocytes	1.41292
Hemoglobin	1.080261
LDH	1.307354
Fibrinogen	1.052542
INR	1.468524
aPTT	1.076211
Blast peripheral blood	1.298846
FAB	1.0814
ELN	1.027841
Type of therapy nominal	1.484999

Table 5.2 VIF values for attributes

The final stage of the preprocessing phase was the selection of features. This was achieved through the application of univariate logistic regression to each feature, with the objective of calculating their respective p-values. The p-value provides an indication of the probability of obtaining a result that is as extreme or more extreme than the observed result when a model that embodies the null hypothesis is employed. In other words, it offers insight into the likelihood that the observed result is a consequence of a change, rather than an intrinsic relationship. To calculate the p-value, it is first necessary to determine the standard error using the following

formula:

$$SE(\hat{\beta}_j) = \sqrt{\text{Var}(\hat{\beta}_j)}$$

Where:

$$\text{Var}(\hat{\beta}_j) = \sigma^2 (X^T X)^{-1}_{jj}$$

The standard error is then employed in conjunction with the estimated parameter to calculate the t-statistic. Based on the t-statistic and degrees of freedom, an appropriate t-distribution may be utilised to calculate the p-value.

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

The results of the univariate logistic regression, applied to all parameters, are presented in the following table:

Variable	P-Value
ECOG_PS	0.006578
Sex	0.023644
Type of therapy	0.026942
Prior history of thrombotic events	0.03687
HCT_CI	0.038144
INR	0.042885
Age	0.135908
LDH	0.166287
Blast peripheral blood	0.177491
AL.ANI.new	0.183525
aPTT	0.211367
HTA	0.254229
Fibrinogen	0.270046
Leukocytes	0.322028
Thrombocytes	0.365618
Smoker	0.429092
Antiplatelet therapy	0.500308
Hemoglobin	0.541803
DM	0.557673
BMI	0.565671

ELN	0.605832
FAB	0.887249

Table 5.3 P-values for attributes

It can be observed that only six parameters have a p-value that is less than 0.05, which is the standard threshold. The aforementioned parameters are ECOG_PS, the patient's sex, the type of therapy, previous thrombosis, HCT_CI and INR. Furthermore, forward and backward selection were conducted, which also demonstrated that utilising more than six attributes has a detrimental impact on performance. To illustrate, we employed Naive Bayes with a forward selection process and these six parameters as the basis. The process augmented the model with three additional parameters: 'LDH', 'Procentat.blasta.krv', 'aPTT'. However, the model with the augmented parameters exhibited inferior performance compared to the original model, with an AUC score of 72.24% compared to 73.49%. A similar pattern was observed with Logistic Regression and K-Nearest Neighbours.

A further advantage of a reduced number of parameters is that the model is more comprehensible and simpler to utilise. Greater explicitness facilitates a more accurate understanding of the parameters that are responsible for positive diagnoses or positive predictions in general. Additionally, the reduction in parameters facilitates the development of a service that is more readily applicable in practice, as medical professionals would be required to collect less data.

6. Results and discussion

The following algorithms were employed in the training process: logistic regression, naive Bayes, k-nearest neighbours, SVM, random forest and neural networks. The SVM exhibited severe underfitting, while the Random Forest demonstrated overfitting to an equivalent degree. Consequently, we ceased utilising these models. The remaining models were trained in three distinct ways: with the default parameters, with a grid search, and with a random search. Models that are capable of accepting weighted examples were trained with those. The reported results are based on the evaluation of the model on previously unseen test data.

The following table presents the performance metrics for the various algorithms.

Algorithms	Logistic regression	K-Nearest Neighbors	MLP	Naive Bayes	Classical Logistic
AUC score	0.734	0.581	0.707	0.716	0.699
Accuracy	0.883	0.814	0.851	0.787	0.883
Sensitivity score	0	0.09	0.136	0.273	0
Specificity score	1	0.9096	0.946	0.855	1
F1 score	0	0.1026	0.177	0.231	0

Table 6.1 Metrics for all the algorithms used.

In this study, the term "classical logistic regression" is used to refer to the logistic regression employed by Mitrovic et al. (2024). The researchers utilised the SPSS software, which employs the Newton-Raphson method without regularisation. This algorithm demonstrated superior performance in all models except k-NN, which exhibited a C score of 85.54% on the training set but only 58.1% on the test set. This suggests that the k-NN model may be prone to overfitting.

The specificity and sensitivity scores for both the classical and logistic regression models are both 1, indicating that the algorithm has predicted that no patients will develop thrombosis. This is a highly misleading result that could potentially cause significant harm. The results were obtained by classifying all instances with an output probability greater than 0.2 as positive. However, this threshold could be modified to establish an optimal balance between positive and negative predictions. The underlying power of the model, as measured by the AUC plot, is not subject to change. The subsequent image depicts the ROC curve:

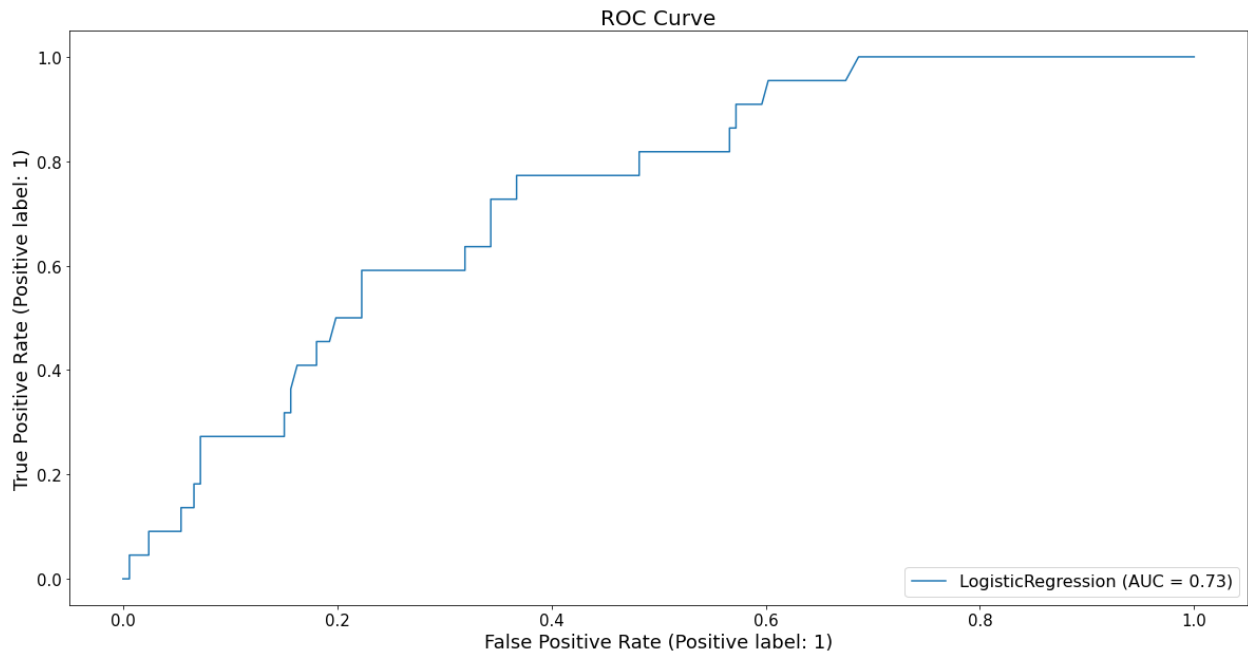


Figure 6.1 Logistic Regression ROC Curve

In this instance, it would be appropriate to select a threshold value that corresponds to one of the identified peaks, given that the values in question are Pareto-dominant in comparison to those that do not represent peaks. The subsequent plot illustrates the impact of varying the threshold on the metrics.

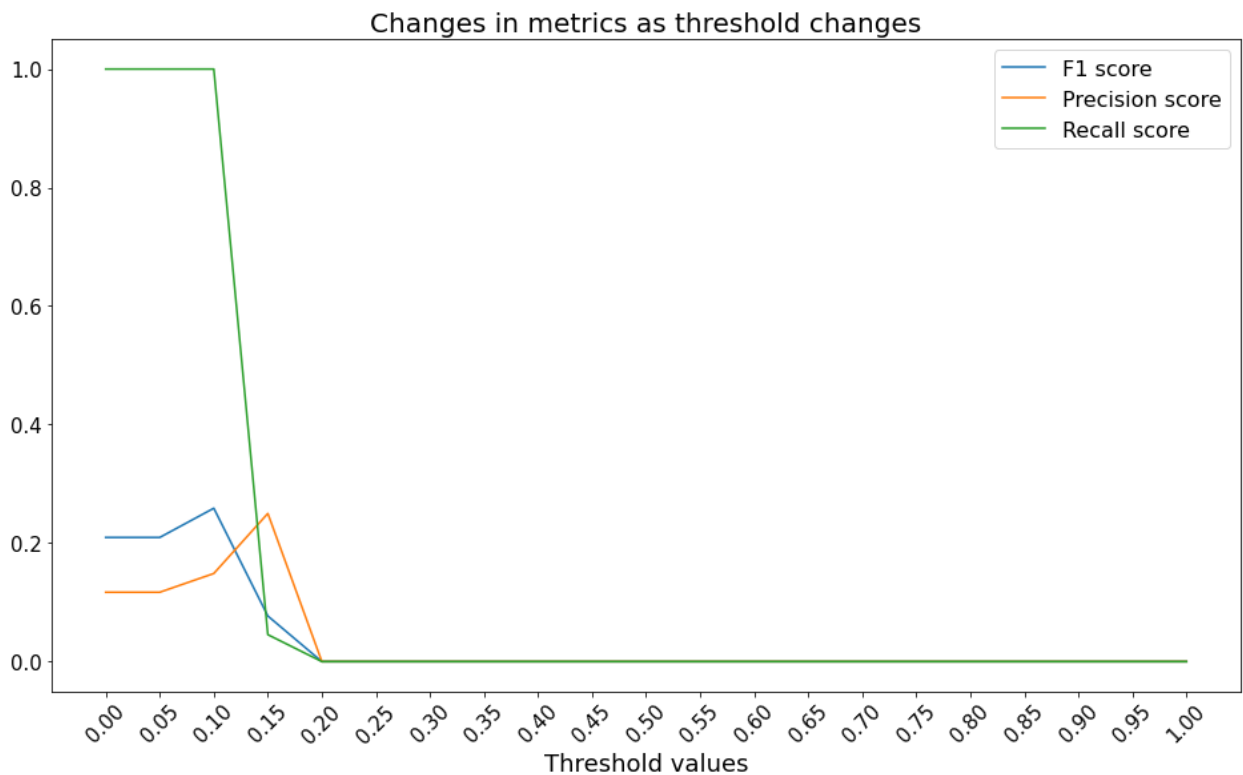


Figure 6.2 Changes in metrics as threshold changes

One of the reasons for this phenomenon is the presence of a considerable imbalance in the number of positive and negative examples, with nine times as many negative examples as positive ones. One potential solution is the utilisation of weighted logistic regression, which

assigns greater weight to the smaller subset and penalises the algorithm more severely when it makes errors on it. The table illustrates the outcomes of this approach. Both models exhibit a perfect sensitivity score and a low specificity score, indicating that the majority of instances are classified as positive and only a few as negative, which is the inverse of the previous outcome.

Algorithms	Classical Logistic regression with weights	Logistic Regression with weights
AUC score	0.705	0.713
Accuracy	0.154	0.139
Sensitivity score	1	1
Specificity score	0.04	0.0241
F1 score	0.2167	0.213

Table 6.2 Metrics for weighted regressions

The analysis of feature influence shows the following:

- Men have higher risk of developing thrombosis
- Lower ECOG_PS scores are associated with greater likelihood of developing thrombosis
- Previous thrombosis, although rare, are strongest predictor of developing next thrombosis
- Intensive therapy, which was often applied, is correlated with a higher probability of venous thrombosis
- International normalised ratio (INR) and hematopoietic cell transplantation-specific comorbidity Index (HCT_CI) are lower in patients that are at higher risk of venous thrombosis

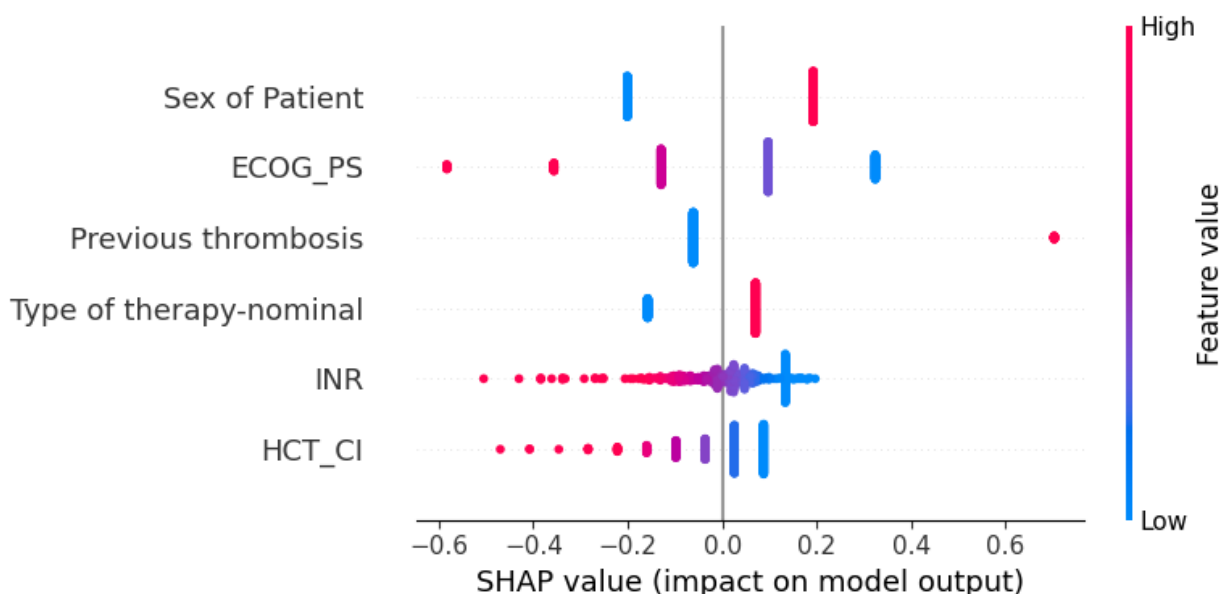


Figure 6.3 The summary plot of SHAP values

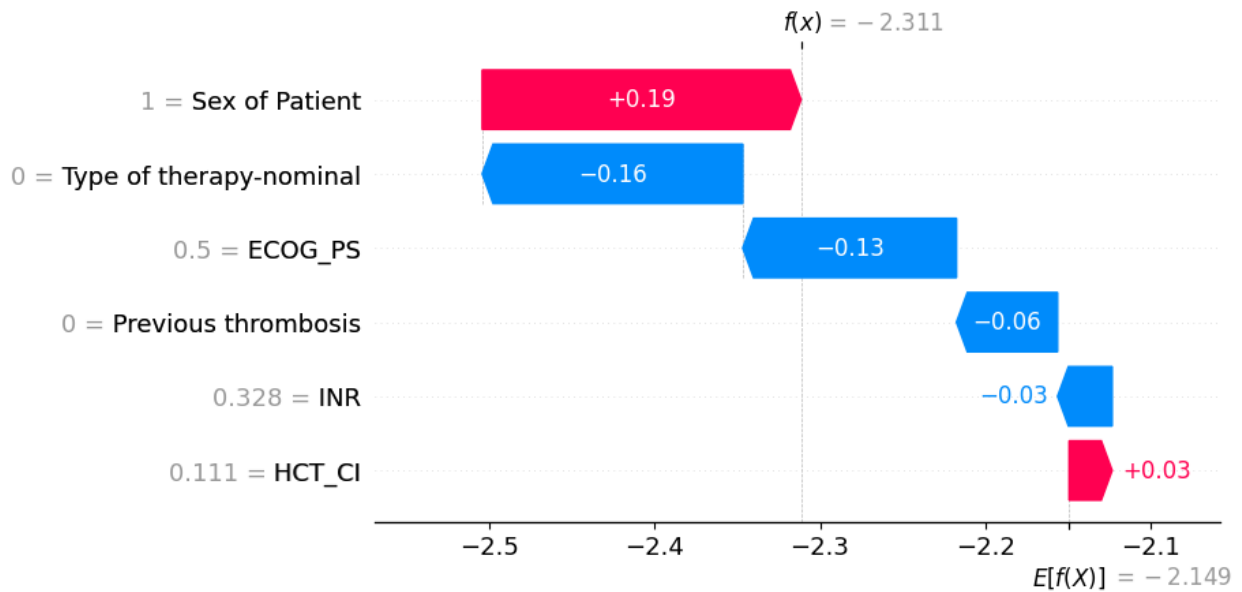


Figure 6.4 The waterfall plot of SHAP values for one example

In the end, logistic regression still proved to be the best-performing model, but unlike in previous work, it utilised an L2 penalty. This regularisation method helped to improve the model's performance by preventing overfitting, making it more robust in handling the dataset. Following logistic regression, the plain Naive Bayes algorithm ranked second in terms of performance, outperforming the neural network. This result is not surprising given the nature of the dataset, which is relatively small and contains only a few informative features. In such cases, Naive Bayes can often excel due to its simplicity and efficiency, as it does not require large amounts of data to perform well. Neural networks, on the other hand, tend to require larger and more complex datasets to fully demonstrate their potential and deliver meaningful results.

In light of the available evidence, it can be concluded that the application's performance has indeed been enhanced through the use of advanced data science methods, although the improvement is limited in scope and was ultimately achieved using relatively simple algorithms. The choice of logistic regression and Naive Bayes highlights the principle that, in many cases, sophisticated machine learning algorithms are not always necessary, especially when dealing with smaller datasets. This work confirms the practical wisdom that, when approaching a data science problem, the best strategy is to begin by training several models, starting from the simple ones and progress from there. This work also underscores a well-known tenet in data science: the performance of a model depends heavily on the size and quality of the dataset.

7. Conclusion

Our research demonstrates how the utilisation of sophisticated data science methodologies can significantly enhance the efficacy of contemporary medical applications, thereby enabling more reliable and precise estimates. Many of the alterations introduced by these sophisticated data science techniques are subtle, and their use does not necessitate extensive technical expertise on the part of healthcare practitioners. For instance, automated algorithms can process large datasets, cleaning and structuring the data without manual intervention, thus enabling even those with limited technical skills to derive meaningful insights. Despite the seeming simplicity of these modifications, they can lead to substantial improvements in the accuracy and reliability of medical predictions, as well as in the efficiency with which healthcare services are delivered.

Furthermore, established industry standards govern the presentation and interpretation of findings derived from data-driven methods. These standards must be carefully considered to ensure that results are not only scientifically sound but also transparent and interpretable by clinicians and other researchers. Adhering to these guidelines helps maintain the integrity of the research process, prevents the miscommunication of complex findings.

Our current research highlights the importance of considering multi-parameter metrics that evaluate a model or system based on multiple parameters or factors. More than single metrics such as accuracy can provide a partial picture of model performance, especially when there are trade-offs between different aspects of outcome evaluation. We suggest that the multi-parameter metric can incorporate different performance measures to provide more comprehensive results, the evaluation of which can support better prediction.

However, it is important to note that statistical algorithms and models are not universally applicable to all datasets or medical conditions. Different types of data and different clinical questions may require customised approaches. Researchers are therefore advised to use a range of models during the training and validation process. This diverse approach allows them to identify those models that show the most promising initial results and the greatest potential for refinement and further development.

In conclusion, the strategic application of data science methodologies in medical research and practice has the potential to revolutionise the field, providing more accurate, timely, and personalised insights into patient care. However, careful consideration must be given to model selection, validation, and the proper communication of results to ensure these advancements translate into real-world improvements in healthcare delivery.

References

- [1] Mack, J. (n.d.). **How data science is reshaping health care**. University of San Diego.
- [2] Lungren, M., Yeung, S., & Cho, M. (n.d.). **AI in healthcare specialization**. Coursera. <https://www.coursera.org/specializations/ai-healthcare>
- [3] Lynch, S. (2017, March 11). Andrew Ng: Why AI is the new electricity. Stanford Graduate School of Business.
- [4] Mathaisel, D. F. X. (2023). Data science in a pandemic. **Data Science Journal, 22*(1), 41.*
- [5] Mitrovic, M., Pantic, N., Bukumiric, Z. et al. (2024). Venous thromboembolism in patients with acute myeloid leukaemia: Development of a predictive model. **Thrombosis Journal, 22*(1), 37.*
- [6] Lee, E. J. et al. (2015). Patterns of venous thromboembolism prophylaxis during treatment of acute leukaemia: Results of a North American web-based survey. **Clinical Lymphoma, Myeloma & Leukaemia, 15*(11), 766–770.*
- [7] Andrews, R. J. (2022, August 1). **How Florence Nightingale changed data visualization forever**. **Scientific American**.
- [8] IBM. (n.d.). **What is ETL?** IBM.
- [9] IBM. (n.d.). **What is data science?** IBM.
- [10] Google. (n.d.). **Google Data Analytics Professional Certificate**. Google.
- [11] Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: Survey and practical guide. **Data Mining and Knowledge Discovery, 18*(1), 140–181.*
- [12] Yanes, J. (2020, December 10). **James Lind and scurvy: The first clinical trial in history**. BBVA OpenMind.
- [13] Bhatt, A. (2010). Evolution of clinical research: A history before and beyond James Lind. **Perspectives in Clinical Research, 1*(1), 6–10.*
- [14] Nash, D. (n.d.). A blink in healthcare.
- [15] Gladwell, M. (2005). **Blink: The power of thinking without thinking**. Little, Brown and Company.
- [16] Brownlee, J. (2020, August 20). How to choose a feature selection method for machine learning. **Machine Learning Mastery**.
- [17] Patro, R. (2021, January 25). Cross-validation: K fold vs Monte Carlo: Choosing the right

validation technique. *Towards Data Science*.

[18] Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical statistics for data scientists: 50+ essential concepts using R and Python* (2nd ed.). O'Reilly Media.

[19] Kahneman, D. (2013). *Thinking, fast and slow*. Farrar, Straus and Giroux.

[20] Sanderson, G. (2019, December). *Bayes' theorem*.

[21] IBM. (n.d.). *What is the k-nearest neighbours (KNN) algorithm?* IBM.

[22] Mukid, M. A., et al. (2018). Credit scoring analysis using weighted k-nearest neighbour. *Journal of Physics: Conference Series, 1025*(1), 012114.

[23] Doknić, I. (2021). *Classification of point clouds using machine learning algorithms* (Undergraduate thesis). University of Novi Sad, Faculty of Technical Sciences, Novi Sad.

[24] Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020). *Practical natural language processing: A comprehensive guide to building real-world NLP systems*. O'Reilly Media.

[25] Sarker I. H. (2021). Data science and analytics: An overview from data-driven smart computing, decision-making, and applications perspective. *SN Computer Science, 2*(5), 377.

[26] Bekbolatova, M., Mayer, J., Ong, C. W., & Toma, M. (2024). Transformative potential of AI in healthcare: Definitions, applications, and navigating the ethical landscape and public perspectives. *Healthcare (Basel, Switzerland), 12*(2), 125.

[27] CUSABIO. (n.d.). *Neuron cell markers*. CUSABIO. <https://www.cusabio.com/Cell-Marker/Neuron-Cell.html>

[28] Ayed, N. (2020, May 28). *The dirt on handwashing: The tragic death behind a life-saving act*. CBC. <https://www.cbc.ca/news/health/the-dirt-on-handwashing-1.5587913>

[29] Shah, R. (n.d.). *Introduction to k-nearest neighbours (kNN) algorithm: A powerful supervised machine learning algorithm. Artificial Intelligence in Plain English*.

[30] Wikipedia contributors. (n.d.). Sigmoid function. Wikipedia. https://en.wikipedia.org/wiki/Sigmoid_function

[31] Brownlee, J. (2023, October). How to use ROC curves and precision-recall curves for classification in Python. Machine Learning Mastery.

[32] Wikipedia contributors. (n.d.). Overfitting. Wikipedia. <https://en.wikipedia.org/wiki/Overfitting>

[33] Popović Krneta, M. Z. (2024). Assessing the accuracy of artificial intelligence models in the application of radioactive iodine therapy and personalised treatment for papillary thyroid cancer patients (Doctoral dissertation). University of Belgrade, Faculty of Medicine.




UNIVERZITET U NOVOM SADU • Prirodno-matematički fakultet 21000 Novi Sad, Trg Dositeja Obradovića 3

KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj, RBR:	
Identifikacioni broj, IBR:	
Tip dokumentacije, TD:	Monografska dokumentacija
Tip zapisa, TZ:	Tekstualni štampani materijal
Vrsta rada, VR:	Master rad
Autor, AU:	Ilija Doknić
Mentor, MN:	Prof. dr Živko Bojović, vandredni profesor
Naslov rada, NR:	Komparativna analiza primene klasičnih statističkih metoda, modela mašinskog učenja i neuralnih mreža u predikciji binarnog ishoda
Jezik publikacije, JP:	Engleski
Jezik izvoda, JI:	Engleski
Zemlja publikovanja, ZP:	Republika Srbija
Uže geografsko područje, UGP:	Vojvodina
Godina, GO:	2024.
Izdavač, IZ:	Autorski reprint
Mesto i adresa, MA:	Novi Sad, Departman za matematiku i informatiku, PMF, Trg Dositeja Obradovića 2
Fizički opis rada, FO:	Poglavlja (7), strana (50), literaturnih citata (33), tabela (5), slika (24)
Naučna oblast, NO:	Matematika
Naučna disciplina, ND:	Primenjena matematika

Ključne reči, KR:	Python, Mašinsko učenje, Klasifikacija, Nauka o podacima, Logistička regresija, KNN, Naivni Bajes, SHAP, Medicina, Neuronske mreže
Univerzalna decimalna klasifikacija, UDK:	
Čuva se, ČU:	Biblioteka Departmana za matematiku i informatiku Prirodno-matematičkog fakulteta, u Novom Sadu
Važna napomena, VN:	
Izvod, IZ:	Model predikcije binarnog ishoda se u širem smislu može posmatrati kao važan alat za rešavanje specifičnih problema u mnogim oblastima, kao što je npr. medicina. U praksi se mogu koristiti različiti modeli, kao što su klasične statističke metode, algoritmi mašinskog učenja i neuronske mreže. Predmet ovog istraživanja je primena navedenih modela za predikciju venske tromboze kod pacijenata sa akutnom mijeloidnom leukemijom. Cilj istraživanja je, da se uradi evaluacija performansi pomenutih modela koristeći različite metrike. Primenjene metode će obuhvatiti: klasične statističke metode (logistička regresija), algoritmi mašinskog učenja (SVM, Naivni Bajes, KNN, Decision Tree) i Neuralne mreže. U analizi će se koristiti Python programski jezika sa odgovarajućim bibliotekama. Dobijeni rezultati će se uporedno analizirati, kako bi se došlo do modela koji na optimalan način procenjuje verovatnoću razvitka tromboze. Time se stvaraju uslovi da se obezbedi adekvatna i pravovremena prevencija.
Datum prihvatanja teme od strane NN veća, DP:	
Datum odbrane, DO:	
Članovi komisije, KO: Predsednik: Mentor: Član:	Prof. Dr Dušan Jakovetić, vanredni profesor, PMF, Novi Sad Prof. Dr Živko Bojović, vandredni profesor, PMF Novi Sad Doc. dr Zoran Bukumirić, vandredni profesor, Medicinski fakultet, Beograd

	UNIVERSITY OF NOVI SAD • Faculty of Science 21000 Novi Sad, Trg Dositeja Obradovića 3	
	KEY WORDS DOCUMENTATION	
Accession number, ANO:		
Identification number, INO:		
Document type, DT:	Monograph type	
Type of record, TR:	Textual material, printed	
Contents code, CC:	Master thesis	
Author, AU:	Ilija Doknić	
Mentor, MN:	Prof. dr Živko Bojović	
Title, TL:	Comparative analysis of applied classical statistical methods, models of machine learning and neural networks in the prediction of binary outcome	
Language of text, LT:	English	
Language of abstract, LA:	English	
Country of publication, CP:	Republika Srbija	
Locality of publication, LP:	Vojvodina	
Publication year, PY:	2024.	
Publisher, PU:	Author's reprint	
Publ. place, PP:	Novi Sad, Faculty of Sciences, Department of Mathematics and Informatics, Trg Dositeja Obradovića 2	
Physical description, PD:	Chapters (7), pages (51), references (33), tables (5), figures (24)	
Scientific field, SF:	Mathematics	
Scientific discipline, SD:	Applied Mathematics	
Key words, KW:	Python, Machine learning, Classification, Data science, Logistic regression, KNN, Naïve Bayes, SHAP, Medicine, Neural networks	

Universal decimal classification, UDC:	
Holding data, HD:	Department of Mathematics and Informatics' Library, Faculty of Science, Novi Sad
Note, N:	
Abstract, AB:	<p>Prediction models for binary outcomes can be considered a crucial tool for solving specific problems in various fields, such as medicine. In practice, different models can be used, including classic statistical methods, machine learning algorithms, and neural networks. The subject of this research is the application of the aforementioned models for predicting venous thrombosis in patients with acute myeloid leukemia. The goal of the research is to evaluate the performance of these models using various metrics. The applied methods will include: classic statistical methods (logistic regression), machine learning algorithms (SVM, Naive Bayes, KNN, Decision Tree), and neural networks. The analysis will utilize the Python programming language with appropriate libraries. The obtained results will be comparatively analyzed to determine the model that best estimates the probability of thrombosis development. This creates conditions for ensuring adequate and timely prevention.</p>
Accepted by the Scientific Board on, ASB:	
Defended on, DE:	
Thesis defend board, DB: Chairperson: Mentor: Member:	<p>Prof. Dr Dušan Jakovetić, professor at PMF, Novi Sad Prof. Dr Živko Bojović, professor at PMF, Novi Sad Doc. dr Zoran Bukumirić, professor at PMF, Novi Sad</p>