Delft University of Technology

Prediction of Lymph Node Classification From a PET-CT Scan

Delft University of Technology, Delft, South Holland, 2628CD

Josephine Clercx (4568095)
Casper Kanaar (4534034)
Elske van Leeuwen (4484460)
Femke Schürmann (4727738)
Amadeo Villar (5377447)
Kelly Vos (4434005)

May 30, 2021

Submission date: May 30, 2021
Course: WI4231
Group: -
Supervisors: Dr. D.B.M. Dickerscheid [1]
Dr. N. Parolya [2]
Drs. M. Wilschut [3]

[1] Medical Physician, Albert Schweitzer Hospital
[2] Assistant Professor, Faculty of Electrical Engineering, Mathematics and Computer Science, Department of Statistics, Delft University of Technology.
[3] Innovator, Albert Schweitzer Hospital

Preface

This project was carried out by 6 MSc students from the faculty of Applied Mathematics and Aerospace Engineering at Delft University of Technology. It is carried out as part of the course WI4231 Mathematical Data Science in collaboration with the Albert Schweitzer Hospital.

We would like to thank our supervisors Dr. N. Parolya from Delft University of Technology, and Dr. D.B.M. Dickerscheid and Drs. Wilschut from the Albert Schweitzer Hospital for this unique opportunity to collaborate with the Albert Schweitzer Hospital, and their continuous support throughout this project.

Source code in R and Python exists for all of the created models in a Github repository. This repository is private because it includes patient data provided by the Albert Schweitzer Hospital. In case the reader wants to access this code, he or she can contact us [4] with his/her Github username.

[4] c.kanaar@student.tudelft.nl

Contents

1 Introduction

This case concerns lung cancer. Since this is the most lethal form of cancer [1], it is important to detect the disease early. If something is found, measures have to be taken almost directly in order to determine as soon as possible what the extent of the disease is. It is important to understand whether the patient suffers from the beginning forms of lung cancer or if the cancer has already spread throughout the body.

Lung cancer spreads through the body from the primary lung tumour via the lymph system. As such, it is important as a stage in between to evaluate the lymph nodes. The lymph nodes of interest for this case are the Hilar lymph nodes in the mediastinum (henceforth simply referred to as lymph nodes). It is essential to examine whether these lymph nodes already contain cancer cells when a primary tumour in the lungs is present. If this is the case, the management of the patient and his or her respective treatment changes radically. For this reason, typically when a suspect mass on a CT scan is observed, patients undergo a PET-CT scan. In a PET-CT scan, a patient is administered a pharmaceutical radioactive tracer via an intravenous injection. The radioactive tracer accumulates in the tumour and from there emits radiation which is detected up by the PET-CT scan. A nuclear medicine physician then determines if this is suspect and whether there might be a lung tumour present. Additional to examining the primary tumour, nuclear medicine physicians examine the lymph nodes as well in order to visually assert whether a lymph node is suspected to be benign or malignant. In case of malignant lymph nodes, a biopsy may be required.

The aim of this report is to construct and assess the performance of a series of models that classify a lymph node to be either benign or malignant based on image data of a primary tumour and its corresponding lymph node provided from CT and PET-CT scans. An annotated database of patients is constructed which contains their 3D CT and PET-CT scans. This project is based on the hypothesis that the image data alone contains information about the pathology of a lymph node. This hypothesis originates from the fact that nuclear medicine physicians are able to learn to detect patterns throughout their career that enable them to tell whether a lymph node is benign or malignant with an accuracy of approximately 70% to 80% [1]. This indicates that information about the pathology of the lymph node is contained within the image data. Additional to inspecting the lymph node, it is required to investigate the visual pattern of the primary tumour. That is, nuclear medicine physicians state that one can not determine from just looking at the lymph node only whether it is benign or malignant, but that the shape and intensity of the primary tumour influences their decision on the evaluation of a lymph node. Additional to the aforementioned aim, the models aim to identify statistical patterns in the 3D volumes of lymph nodes and primary tumours that correlate to the pathological outcome. This means that image data is being investigated from a statistical point of view. This adds to the following research question as provided by the Albert Schweitzer Hospital:

"Is it possible to predict if a lymph-node is benign or malignant from the properties of the 3D segmented lymph-node and the primary tumour of the patient?"

[1] ASZ Cases by Dr. D.B.M. Dickerscheid and Drs. Wilschut [Last Accessed 26-05-2021]

Additional sub-research questions as provided by the Albert Schweitzer Hospital are:

- "How much can the predictive performance be increased with respect to the simple thresholding model proposed by the Albert Schweitzer Hospital?"
- "Does the information from the primary tumours help to improve lymph node classification predictions?"
- "Are there differences between the datasets of the two different hospitals?"
- "Does the CT data correlate with the classification of the lymph nodes and if so in what respect?"

The aforementioned aim of the report is realised through the following structure. First, a description and analysis of the data is provided in Chapter 2. Next, the theory behind each model considered for analysis is described in Chapter 3. Following this, the models are applied to the classification problem and their performances are analysed in terms of various measures in Chapter 4. Based on this performance assessment, a conclusion is drawn in Chapter 5 on whether the models are suitable for the classification of a lymph-node being benign or malignant from the properties of the 3D segmented lymph-node and the primary tumour of the patient. Further conclusions are also drawn on the additional sub-research questions. Finally, a discussion is presented in Chapter 6 based on the process and findings of this report.

2 Data description and analysis

The provided data contains information of 145 patients who have been diagnosed with a primary tumour present in the mediastinum. 115 of these patients are treated at the Albert Schweitzer Hospital (ASZ) and 30 patients are treated at the Diakonessen Hospital (DIAK). Additional to the primary tumour data of each patient, data on a single or multiple Hilar lymph nodes is provided. The lymph node data is labelled into two classes: benign and malignant. This means that classification models can be trained based on the image data of the primary tumour and the lymph nodes to assert whether a specific lymph node is benign or malignant. The data of the primary tumour and the lymph node(s) come in the form of 3D image data, and are obtained through different medical imaging techniques, namely a CT and a PET scan (where the PET scan data is obtained through a combined PET-CT scan)[1].

A description and exploration of the CT and PET data are provided in Section 2.1 and Section 2.2, respectively. Next, the data is formatted into features such that a feature design matrix or data frame can be created. This formatting is performed in Section 2.3. Finally, this constructed data frame is analysed in Section 2.4 in order to establish some preliminary results relevant to the construction and assessment of the models.

2.1 CT scan data

When a patient is referred to the hospital by his or her general practitioner because of suspicions of lung disease, a pulmonologist might determine additional tests are needed. In this case, a patient may undergo a CT-scan. As such, lymph node data is provided in the form of CT image data. The CT image data is provided in the form of $144 \times 144 \times 144$-sized `numpy` arrays with voxel intensity. The voxel intensity represents gray-scale intensity of the CT scan image and is therefore in the Hounsfield unit [HU] Kalender (2011). An example of a lymph node as given by a CT scan is shown in Figure 2.1. Note that the four different sub-figures indicate different depth levels of the same lymph node.

2.2 PET-CT scan data

If a suspect mass is detected on a CT scan, a patient might undergo a PET-CT scan. As such, primary tumour data is provided only in the form of PET image data, whereas lymph node data is provided in both the form of PET image data and CT image data. The PET image data is provided in the same format at the CT image data. That is, in $144 \times 144 \times 144$-sized `numpy` arrays with voxel intensity. The PET image data voxels are given in units of [Bq/ml]. They are a measure for the activity concentration of the radioactive F18-FDG tracer within the body. Furthermore, the

[1] ASZ Cases by Dr. D.B.M. Dickerscheid and Drs. Wilschut [Last Accessed 25-05-2021]

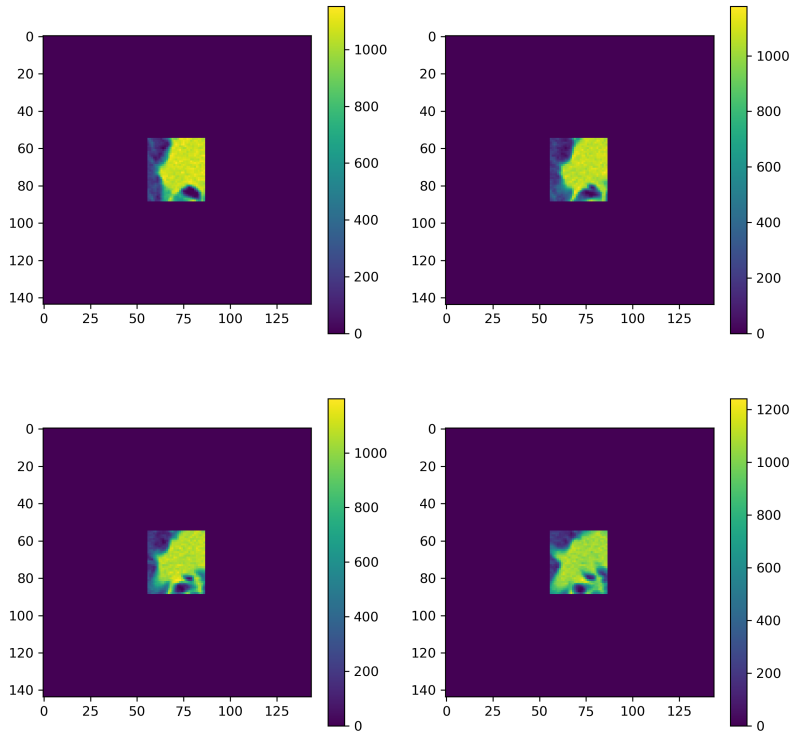


Figure 2.1: Example of a lymph node as given by a CT scan.

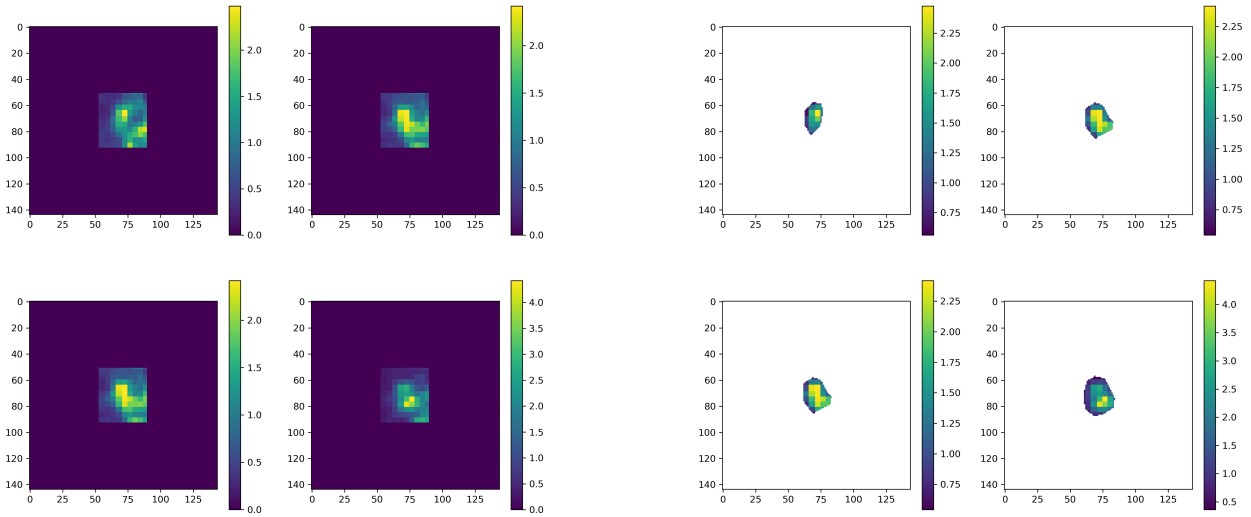
PET image data comes with additional MASK files. When applied to the PET data, these MASK files segment the primary tumour or lymph node from the image. This masking is exemplified by Figure 2.2a and Figure 2.2b, which show that a mask file segments the primary tumour (for this specific example) from the image. The four subfigures are four different depth levels of the same lymph node.

Additional to the PET image data, the patient ID, hospital of treatment, location of the lymph node, patient mass, and SUV factor are provided. The SUV factor is used to normalise the voxels of the PET data. Since the voxels of the PET data are in units of [Bq/ml], they are a measure for the activity concentration of the radioactive F18-FDG tracer within the body. This means that this concentration must be read in context of the patient's mass and time that has passed since administering the dose. The Standardised Uptake Value (SUV) incorporates this. The SUV is indicative of the accumulation of radioactive tracer by comparing the tracer uptake level to that of a completely homogeneous distribution. The SUV factor is given by (2.1)[2].

$$\text{SUV} = \frac{v}{\text{SUV factor}} \tag{2.1}$$

In (2.1), v is the voxel intensity value in [Bq/ml] and the SUV factor is a correction factor based on

[2] `TUDelft_nodeanalysiscase.ipynb` by Dr. D.B.M. Dickerscheid [Last Accessed 25-05-2021]



(a) Unmasked. (b) Masked.

Figure 2.2: Example of a primary tumour as given by a PET-CT scan.

the decay of the F18-FDG tracer and the patient mass, as given by (2.2) [3].

$$\text{SUV factor} = 1000 \cdot 0.97 \cdot \frac{A_0 \left(\frac{1}{2}\right)^{\frac{\Delta t}{T_{1/2}}}}{m} \tag{2.2}$$

In (2.2), A_0 is the administered activity in units of [MBq], Δt is the time interval between administering the activity dose A_0 and the acquisition of the image, m is the patient mass in [kg] and $T_{1/2}$ is the half life time of the F18-FDG tracer in units of [s]. Note that the voxel intensities of Figure 2.2 have been normalized by the SUV factor.

2.3 Formatting data features

The models described in Chapter 3 classify lymph nodes to be either benign or malignant based on features of the primary tumours and lymph nodes. As such, it is desirable to extract these features from the data described in Section 2.1 and Section 2.2 and to create a data frame containing these relevant features which can serve as a design matrix for the upcoming models. These features included in the data frame are given as follows:

- Patient ID (ID);
- Hospital of Treatment (HOS);

[3] `TUDelft_nodeanalysiscase.ipynb` by Dr. D.B.M. Dickerscheid [Last Accessed 25-05-2021]

- Primary Tumour Volume (PV);
- Primary Tumour Mean SUV (PSUV);
- Lymph Node Volume (LV);
- Lymph Node Mean SUV (LSUV);
- Lymph Node Mean HU Value (LHU).

The label corresponding to these features is then the pathological result of the lymph node (benign or malignant):

- Pathological Result.

It is expected that accurate classification results can be obtained based on these features. Note that some models may not use each of these features in their respective classification. The reason for including each of the features into the data frame is as follows:

Patient ID (ID) Since multiple lymph nodes can correspond to the same patient and thereby the same primary tumour, it is important to keep track of which primary tumour belongs to which lymph node. For this reason, patient ID is included in the data frame. In this manner, multiple rows in the data frame can correspond to a single patient where each row contains the same primary tumour data, but contains different lymph node data.

Hospital of Treatment (HOS) As mentioned in one of the sub-research questions in Chapter 1, it may be interesting to assess whether there is a difference in data between the data provided by the Albert Schweitzer Hospital and the Diakonessen Hospital. For this reason, the hospital at which the patient was treated has been incorporated in the dataset. In this way, it can easily be retrieved to which hospital a certain data point corresponds, leading to more convenient analysis of results.

Primary Tumour Volume (PV) The size of the primary tumour is expected to correlate to the pathological result of the lymph nodes. For this reason, the primary tumour volume has been included in the data frame as a measure of size. The volume of primary tumour is measured in units of number of voxels after applying the masks to the 3D PET-CT scan data.

Primary Tumour Mean SUV (PSUV) Since the intensity of the primary tumour in terms of SUV is also expected to correlate to the pathological result of the lymph nodes, it has been included in the data frame. The mean SUV value of the primary tumour was computed by first dividing the raw PET-CT voxel data by the respective SUV factor as given by (2.1), then applying the mask to the data to obtain only the relevant segmented voxels, and finally taking the mean of the SUV factor normalized voxel intensities.

Lymph Node Volume (LV) The volume of the lymph node is incorporated in the data frame since it is expected to correlate strongly its the pathological outcome. Similar to PV, lymph node volume is measured in units of number of voxels after applying the masks to the 3D PET-CT scan data.

Lymph Node Mean SUV (LSUV) Since the lymph node mean SUV is expected to correlate with its pathological outcome, it is included in the data frame. This mean value is computed similarly to the mean SUV of the primary tumour. First the voxel intensity is divided by the respective SUV factor as given by (2.1), next the mask is applied to segment the lymph node's relevant voxels, and finally the mean over the SUV factor normalized voxel intensities is taken.

Lymph Node Mean HU Value (LHU) In order to also incorporate the CT scan data in the data frame, the mean HU value of the lymph node's is included. The mean HU value of the lymph node relates to the intensity of the lymph nodes which is expected to correlate with the pathological outcome as described earlier. The lymph node mean HU value is computed by first applying the mask files to the CT scan data such that only segmented lymph nodes are left, and then taking the mean of the voxel intensity. Note that contrary to the PET-CT scan data, no additional pre-processing is required for the voxels of the CT scan data (no division by SUV factor).

2.4 Data analysis

In this section, the data frame constructed in Section 2.3 is analysed in order attain an understanding of exactly what data the models presented in Chapter 3 have to work with. First, the data is summarised in Subsection 2.4.1. Next, a univariate analysis applied on each of the variables in Subsection 2.4.2 in order to examine each of the variables separately. Following this, a bivariate analysis is applied to the variables in Subsection 2.4.3 in order to understand the correlation between all variables. Continuing, an outlier analysis is performed in Subsection 2.4.4, followed by a hospital analysis in Subsection 2.4.5. Finally, the data is split into a train and test set in Subsection 2.4.6.

2.4.1 Summary of the data

To summarise the data frame as constructed in Section 2.3, the `str` function in R is used:

```
    >str(data)
'data.frame': 504 obs. of  8 variables:
 $ ID                  : int  102 102 102 102 102 102 104 104 107 107 ...
 $ HOS                 : Factor w/ 2 levels "ASZ","DIAK": 1 1 1 1 1 1 1 1 1 1 ...
 $ PV                  : int  3473 3473 3473 3473 3473 3473 10728 10728 34291 34291 ...
 $ PSUV                : num  1.77 1.77 1.77 1.77 1.77 ...
```

```
$ LV                  : int  1583 943 1664 758 171 462 11482 4667 28977 2013 ...
$ LSUV                : num  2.45 2.64 2.65 2.81 3.02 ...
$ LHU                 : num  900 786 984 1044 931 ...
$ Pathological.Result : Factor w/ 2 levels "Benign","Malignant": 1 1 1 1 1 1 1 1 2 2 ...
```

The data frame contains 3 integer variables (PV, LV and ID), 2 factor variables (HOS and Pathological Result), and 3 float variables (PSUV, LSUV and LHU). All values are positive. First we develop a descriptive analysis where we collect the five number summary statistics of our dataset using the `summary` command.

```
   >summary(data)
        ID     HOS         PV                    PSUV
Min.   :  2   ASZ :406   Min.   :   772        Min.   : 0.9688
1st Qu.:133   DIAK: 98   1st Qu.: 10240        1st Qu.: 3.1859
Median :229              Median : 27464        Median : 4.4583
Mean   :288              Mean   : 56618        Mean   : 5.0011
3rd Qu.:348              3rd Qu.: 68690        3rd Qu.: 6.2448
Max.   :916              Max.   :395393        Max.   :16.4618

LV                LSUV                    LHU
Min.   :  147     Min.   : 1.616          Min.   : 187.8
1st Qu.: 1184     1st Qu.: 3.165          1st Qu.: 908.7
Median : 2794     Median : 4.288          Median : 983.2
Mean   : 7305     Mean   : 4.614          Mean   : 936.2
3rd Qu.: 8998     3rd Qu.: 5.582          3rd Qu.:1027.1
Max.   :97120     Max.   :12.940          Max.   :1202.5
Pathological.Result
Benign   :117
Malignant:387
```

With this function, we have obtained the minimum and maximum, the first and third quartile, the mean and the median values of the numerical/integer values. Furthermore, the summary shows a frequency of both the factor variables. Note that, the results for the variable patient ID are not relevant for the pathological outcome, but are simply a measure to keep track of which lymph node belongs to which tumour in the data frame. As such, ID is discarded from the analysis henceforth.

2.4.2 Univariate analysis

Next, we apply an univariate analysis which consists of examining all the variables separately. We start with analysing the pathological result variable. For this we use a bar plot as given by Figure 2.3 to visualise the frequencies and a table of frequency given by Table 2.1a to count the percentage of the value that are in the different categories.

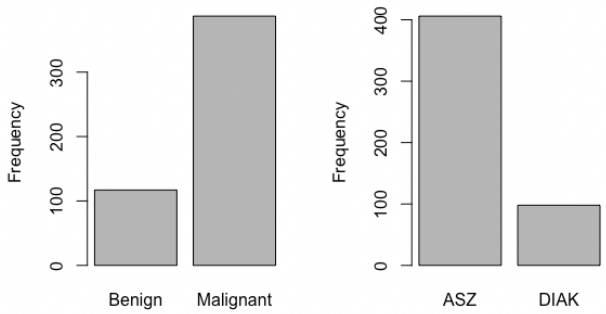


Figure 2.3: Frequency plots of the pathological results and the two hospitals

Table 2.1: Frequency tables

(a) Lymph nodes' pathological results

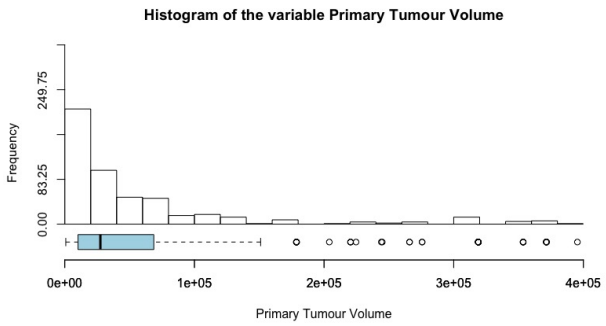
	Benign	**Malignant**
Number	117	387
Percentage	23%	77%

(b) Patients' hospital of treatment

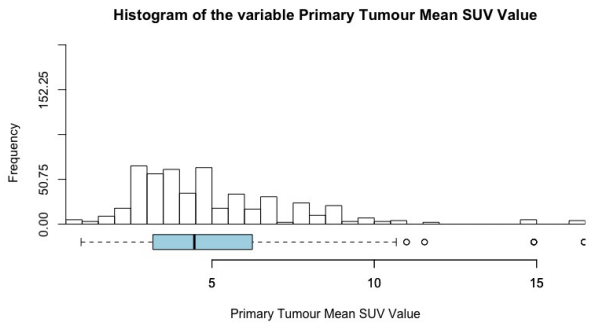
	ASZ	**DIAK**
Number	406	98
Percentage	81%	19%

Analysing the left plot in Figure 2.3 and Table 2.1a, it can be seen see that the dataset contains many more malignant results than benign results. To be more specific, 117 (23%) cases are benign whereas 387 (77%) cases are malignant. In other words, the data is biased. For the hospitals, it can be seen in the right plot of Figure 2.3 and Table 2.1b that most data comes from the ASZ hospital (406 cases, 81%) whereas only 98 (19%) cases come from the DIAK hospital. Also, here we see that the data is biased. Note since the hospital does not have any effect on the chance of lymph node being either benign or malignant, it is excluded from analysis in the models described in Chapter 3. Instead, it is included in the data frame to investigate possible differences between the data from the Albert Schweitzer Hospital and the Diakonessen Hospital.

Knowing that the data is biased, we are going to follow the stratified sampling method for splitting the data into the test and training set. This is further described in Subsection 2.4.6. Next, we proceed with the other variables. For this we create boxplots and histograms for the variables, as shown by Figure 2.4 up to including Figure 2.6.

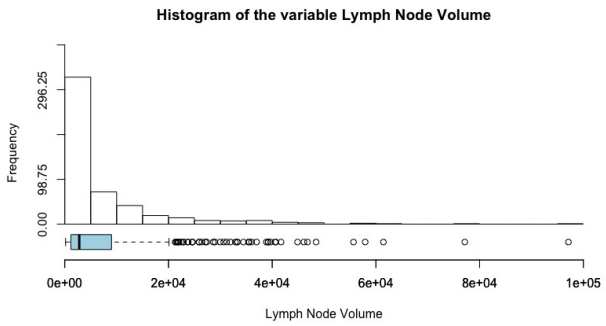


(a) Histogram and boxplot for PV

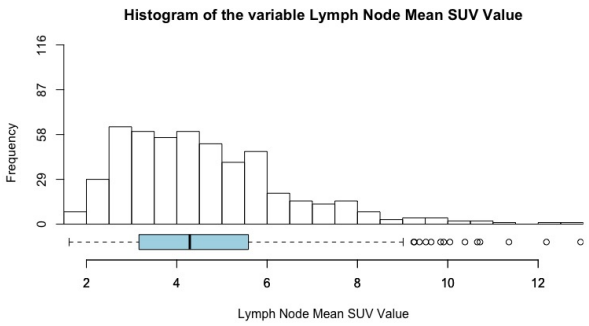


(b) Histogram and boxplot for PSUV

Figure 2.4: Histograms and boxplots for PV and PSUV



(a) Histogram and boxplot for LV



(b) Histogram and boxplot for LSUV

Figure 2.5: Histogram and boxplot for LV and LSUV

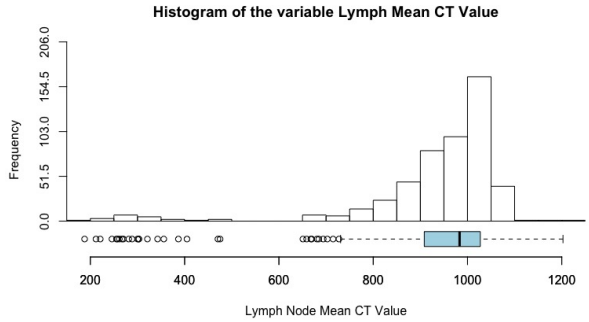


Figure 2.6: Histogram and boxplot for LHU.

We see that the two volume variables have a right skewness in their distribution, the LHU variable has a left skewness. The histograms of the two SUV variables have a wider range than the other variables, most of these values are grouped at the left side of the graph. These phenomenons indicated that the variables have quite some outliers. The data of the lymph node variables (LV, LSUV, and LHU) contain the most outliers.

2.4.3 Bivariate analysis

A bivariate analysis is used to understand the relationship between all variables. The bivariate analysis is summarised by the correlation matrix in Figure 2.7.

In Figure 2.7, the positive and negative correlation between variables can be observed. There is a positive correlation of 0.22 between the PV and the PSUV. Also, there is a positive correlation between the LSUV and the variables PSUV (0.41) and the LV (0.36). The negative correlations are quite minimal, there is a negative correlation of -0.15 between the LV and the LHU. The other negative correlations are negligible.

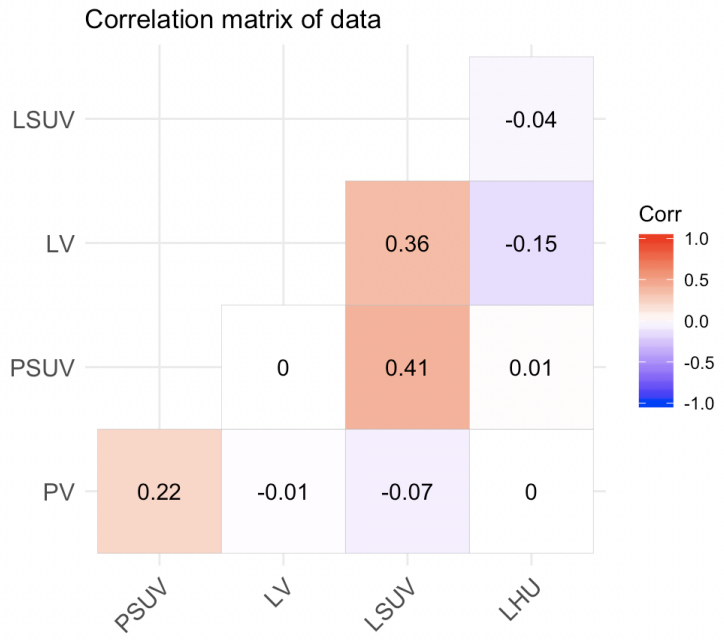


Figure 2.7: Analysis of the correlation between variables

2.4.4 Outlier analysis

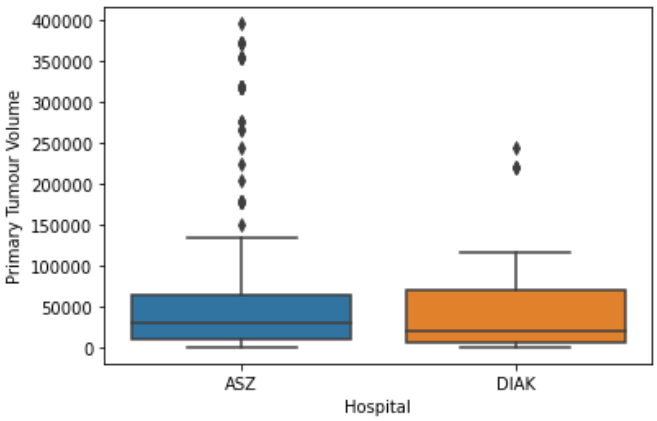
An outlier analysis is performed next. In order to do this, an outlier needs to be defined. A datapoint is an outlier when it is 1.5 times the inter quartile range away from the inter quartile range.

		Number of outliers
1	PV	44
2	PSUV	13
3	LV	48
4	LSUV	14
5	LHU	32

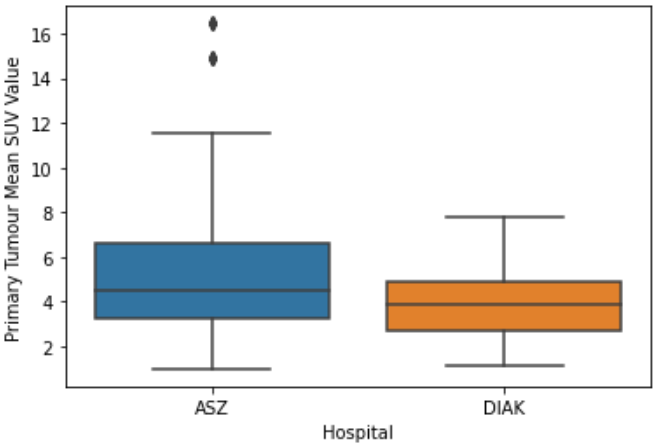
All variables have outliers. In order to make sure that important information is not removed from the data frame, all outliers are accepted and none are removed.

2.4.5 Hospital analysis

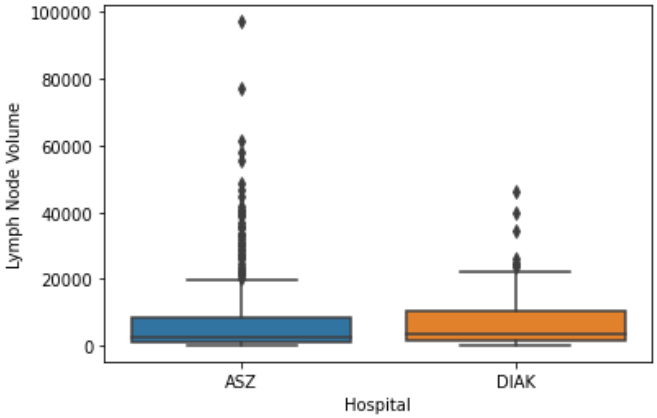
One of the sub-research questions stated in Chapter 1 is directed at whether there are differences between the data provided from the Albert Schweitzer Hospital, and the Diakonessen Hospital. To this purpose, boxplots are made in such that the difference in relevant statistics can be visualised. These boxplots are given in Figure 2.8.



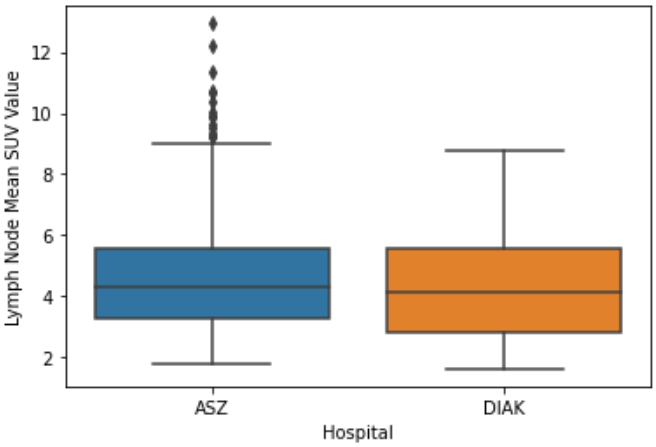
(a) Primary tumour volume.



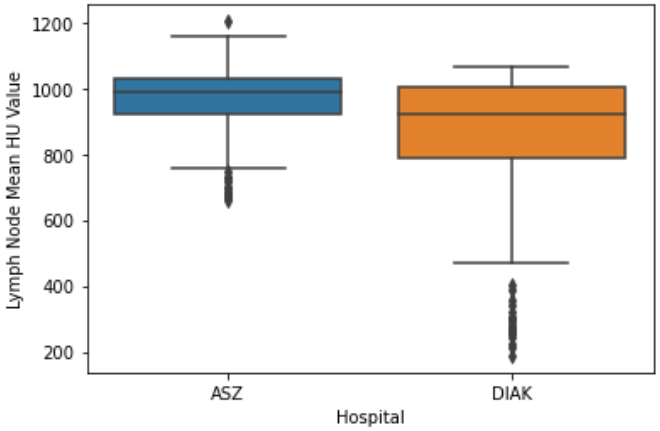
(b) Primary tumour mean SUV value.



(c) Lymph node volume.



(d) Lymph node mean SUV value.



(e) Lymph node mean HU value.

Figure 2.8: Boxplots of the predictor variables for the two different hospitals (Albert Schweitzer Hospital and Diakonessen Hospital).

From Figure 2.8, it can be seen that there is not a significant difference between the data of the Albert Schweitzer Hospital and the Diakonessen Hospital. The outliers are a little bit different, but in general the boxplots are very similar. Especially the means are close, and a small difference in

mean could be explained from the fact that the Diakonessen data is considerably smaller than the Albert Schweitzer data set.

2.4.6 Generating train and test data

In order to effectively assess the performance of the models in Chapter 4, the assessments need to be conducted with respect to data which the models have not "seen" before. After all, when applied in practice the models will have to classify a lymph node which has not been included in the train set. In the following parts we will correctly arrange the dataset for prediction. The data is split into a 70% train set and a 30% test/validation set, where the choice of ratio between train and test data has been chosen based on common ratios. The 30% test set is not used to train the model but rather to assess the performance. Since the data is biased as explained in Subsection 2.4.1, a stratified sampling technique is used to split the train and test data. The result of this is given in Table 2.2.

Table 2.2: Train and test data.

	Benign	**Malignant**
Train set	82	270
Test set	35	117

From Table 2.2, it can be seen that the train set contains more samples. However, the stratified sampling techniques makes sure that the ratio of benign to malignant cases is the same for the train set and the test set (ca. 23%). In Chapter 4, the same test and train set is used for the training and evaluation of each model.

3 Model descriptions

Before applying the models in practice on the provided data set, it is important to develop an understanding of their workings. In this manner, advantages and limitations of the models can more easily be traced and understood. For this reason, this chapter features theoretical descriptions of the models. First, the thresholding model provided by the Albert Schweitzer Hospital is discussed in Section 3.1. Next, the logistic regression model is discussed in Section 3.2. Continuing, the support vector machine model is discussed in Section 3.3. The final model is the decision tree model and is presented in Section 3.4. Following the model descriptions, Section 3.5 finalises by listing the performance measures used for model assessment in Chapter 4.

3.1 Thresholding

In this section we are going to study the thresholding model, proposed by the Albert Schweitzer Hospital [1]. This model is the simplest of the models that we will develop during this project, since it only takes takes into account the LSUV variable studied in Section 2.3. This model is based on establishing a threshold level, so that lymph nodes with a value higher than the threshold level are considered malignant and those with a lower value are considered benign:

$$\begin{cases} \text{if LSUV} \geq \text{Threshold}_{\text{level}} \rightarrow Y = 1 \text{ (malignant lymph node)} \\ \text{if LSUV} < \text{Threshold}_{\text{level}} \rightarrow Y = 0 \text{ (benign lymph node)} \end{cases} \tag{3.1}$$

The hypothesis for establishing a threshold value based on the LSUV variable is based on the observation that the malignant LSUV values are in general higher than the benign LSUV values, as shown by Figure 3.1. Furthermore, the variable analysis for each other model evaluated in Chapter 4 indicates that LSUV is the most significant variable.

The threshold level is established based on the train set. A range of threshold values in [min(LSUV), max(LSUV)] is considered. The threshold value that results in the maximum accuracy on the train set is then used to evaluate the test set. The results obtained with this model are presented in Section 4.1.

[1] `TUDelft_nodeanalysiscase.ipynb` by Dr. D.B.M. Dickerscheid [Last Accessed 25-05-2021]

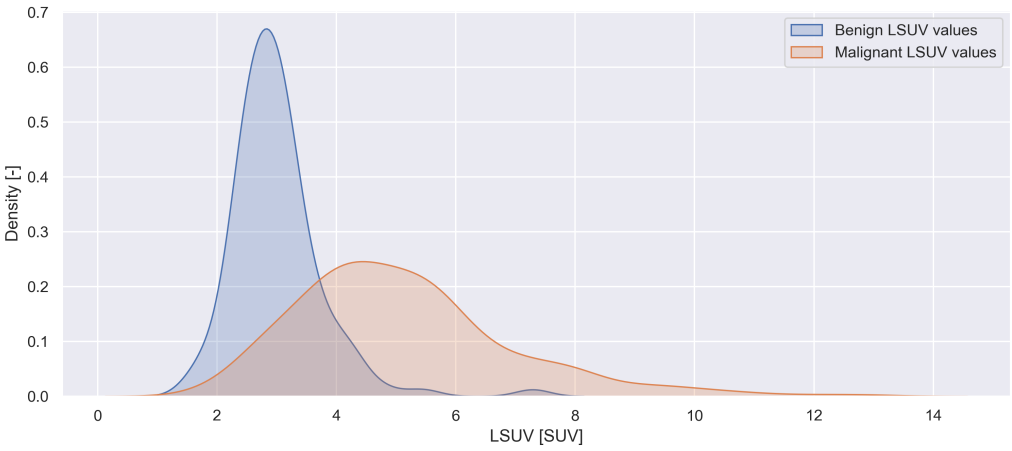


Figure 3.1: KDE of LSUV values for benign and malignant lymph nodes.

3.2 Logistic regression

After having studied the simpler model of thresholding in Section 3.1, we move on to study more accurate and complex statistical models. In this section we are going to study the so-called logistic regression. Logistic regression is a type of regression that uses a logistic function to model a categorical dependent variable Hastie et al. (2001). In this section we will focus on the case where the random variable is binary, but this model can also be extended to the multivariate case. This type of model is widely used in the field of machine learning, and has many applications in classification problems Hastie et al. (2001). Among them we highlight the field of medical testing, in which we can include the problem we are dealing with in this project. Subsection 3.2.1 starts by discussing the setup of the model. Next, Subsection 3.2.2 describes the estimator for the regression coefficients. Following this, the model's prediction criterion is presented in Subsection 3.2.3. Continuing, Subsection 3.2.4 lists the performance that will be assessed to evaluate the performance of the model. Finally, Subsection 3.2.5 closes off with some additional remarks regarding the logistic regression model. The results obtained with this model are presented in Section 4.2

3.2.1 Setup

The setup is as follows: we have a random variable Y, which models the pathological outcome of a lymph node. This variable Y has 2 possible outcomes: benign or malignant. To lighten notation, we assign to each of these outcomes the labels 0 and 1 respectively, so that we can understand Y as a Bernoulli random variable $\rightarrow Y \sim Ber(p)$.

Our fundamental objective is to study how we can obtain an estimator of p for a given lymph node. This parameter p is conditioned by a series of independent random variables $X_1, \ldots, X_n$, which we call predictors. In our case, these predictors are the data features that we studied in Section 2.3, and they can be either continuous or discrete random variables. The problem therefore turns to

estimating the posterior probability $p = \mathbb{P}(Y = 1 \mid X_1 = x_1, \ldots, X_n = x_n)$ given determined data features. As we are dealing with a binary problem, the probability that the lymph node is benign is $1 - p$.

The logistic regression model is based on the assumption that the logarithm of the odds (ratio between the likelihood of the lymph nodes being malignant and benign) depends linearly on the predictors. This is expressed by

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n, \tag{3.2}$$

or by clearing the value of p in equation (3.2), which gives the equations

$$p = \mathbb{P}(Y = 1 \mid X_1 = x_1, \ldots, X_n = x_n) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}}, \tag{3.3}$$

$$1 - p = \mathbb{P}(Y = 0 \mid X_1 = x_1, \ldots, X_n = x_n) = \frac{1}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}}. \tag{3.4}$$

The regression problem is then reduced to obtaining the parameters $\beta_0, \beta_1, \ldots, \beta_n$ that characterise the previous linear dependence. This logistic function distinguishes the logistic regression model, but there are analogous models which use different sigmoid function, such as the probit model, or models taking logarithm in different basis. Although in this project we will only focus on logistic/logit model.

3.2.2 Regression coefficients

The regression coefficients are usually estimated using maximum likelihood estimation. Given our train data set $\{Y_1, \ldots, Y_m\}$, with predictors $\{X_{j,1}, \ldots, X_{j,n}\}$ with $j = 1, \ldots, m$, the estimator for the regression coefficients are those which maximise the likelihood given by

$$L(\beta_0, \ldots, \beta_n \mid y_1, \ldots, y_m) = \prod_{j=1}^{m} \left(\frac{e^{\beta_0+\beta_1 x_{j,1}+\cdots+\beta_n x_{j,n}}}{1 + e^{\beta_0+\beta_1 x_{j,1}+\cdots+\beta_n x_{j,n}}}\right)^{y_j} \left(\frac{1}{1 + e^{\beta_0+\beta_1 x_{j,1}+\cdots+\beta_n x_{j,n}}}\right)^{1-y_j}. \tag{3.5}$$

Since we can not find a closed-form expression for the coefficients in (3.5), we use the iterative Fisher's scoring algorithm Longford (1987) to find these estimators.

The estimators can be interpret as follows:

- $\hat{\beta}_0$ is the y-intercept. By exponentiating it, we have the odds of the event that the lymph node is malignant, $Y = 1$, when the predictors $X_1 = 0, \ldots, X_n = 0$. In our case we will check in Chapter 4 that this value is very small, since within our predictors we have volume of the tumour and SUV values. Then the chances of a lymph node being malignant when there is no tumour (tumour volume is 0) are logically 0.
- $\hat{\beta}_j$ is the linear coefficient associated to X_j. Increasing X_j by 1 unit, the odds that the lymph node is malignant will increase/decrease by a factor $e^{\hat{\beta}_j}$ depending on whether $\hat{\beta}_j > 0$ or $\hat{\beta}_j < 0$

respectively. Therefore the sign of the estimator will indicate if the associated predictor is directly proportional or inversely proportional. Indeed,

$$OR = \frac{\text{odds}(x_j + 1)}{\text{odds}(x_j)} = \frac{e^{\hat{\beta}_0 + \cdots + \hat{\beta}_j x_j + \hat{\beta}_j + \cdots + \hat{\beta}_n x_n}}{e^{\hat{\beta}_0 + \cdots + \hat{\beta}_j x_j + \cdots + \hat{\beta}_n x_n}} = e^{\hat{\beta}_j}. \tag{3.6}$$

3.2.3 Prediction criterion

Once $\hat{\beta}_0, \ldots, \hat{\beta}_n$ are estimated, the model is complete and we can easily compute the probability that a given lymph node is malignant. In practice, to make predictions about whether $Y \mid X = 0$ or $Y \mid X = 1$ we need a criterion based on these calculated probabilities. A generally used criterion is to set a significant level α, such that

$$\begin{cases} \text{if } p \geq \alpha \to Y = 1 \text{ (malignant lymph node)} \\ \text{if } p < \alpha \to Y = 0 \text{ (benign lymph node)} \end{cases} \tag{3.7}$$

Normally $\alpha = 0.5$ is taken, but we will see later that in our case, since our data is biased (Figure 2.3), the best predictions are obtained for $\alpha \approx 0.65$. Other possible choice would be to use a mathematical software to generate random samples of a Bernoulli random variable given the probabilities computed. This is discussed in Section 4.2.

3.2.4 Goodness-of-fit tests

After having estimated the coefficients and established a criterion for making predictions, we ask ourselves how well our model performs. To answer this question, we will study a series of goodness of fit tests for our model. Among them we find fit tests that are also applicable to the decision trees model and the support vector machine model. In this project we will deal with confusion matrices, ROC curves and area under these curves, which will allow us to make comparison between the different models. These measures are further explained in Section 3.5.

On the other hand, there are fit tests that can only be applied to the logistic regression model. Among them the likelihood ratio test and Deviance tests are quite important Hastie et al. (2001). These tests are used when a "saturated" model (a model with a theoretically perfect fit) is available. In these cases the deviance, defined as

$$D_{\text{fitted}} = -2 \log \left(\frac{L_{\text{fitted}}}{L_{\text{saturated}}} \right). \tag{3.8}$$

allows us to make statements about the goodness of our model. The smaller the deviance, the closer is the fitted model to the saturated model, and hence the better its performance. On the contrary large values of D indicate lack of fit. Note that in (3.8), L denotes likelihood.

However, in practice it is very difficult if not impossible to obtain a saturated model. It is for this reason that we introduce the null model instead. The null model is the model in which only the

intercept is taken into account, i.e. the case in which there are no predictors. The null deviance is therefore

$$D_{\text{null}} = -2\log\left(\frac{L_{\text{null}}}{L_{\text{saturated}}}\right). \tag{3.9}$$

By subtracting the above expressions, we obtain an independent statistic of the saturated model, that we can apply to our model

$$D_{\text{null}} - D_{\text{fitted}} = -2\log\left(\frac{L_{\text{null}}}{L_{\text{fitted}}}\right). \tag{3.10}$$

The smaller this previous statistic is, the closer is the fitted model to the null model, and hence the worse its performance. Therefore, we will be interested in obtaining high values for this statistic.

Closely related to (3.10) are the test based on measuring the pseudo-R^2 statistic. This quantity is the analogue to the R^2 statistic used in linear regression, but applied in the logistic regression frame. During this project we will use the so-called McFadden R^2_{McF} index Domencich & McFadden (1975), given by

$$R^2_{\text{McF}} = 1 - \frac{\log(L_{\text{fitted}})}{\log(L_{\text{null}})}. \tag{3.11}$$

This index is always between 0 and 1. When our model has very little predictive ability, the likelihood value of the fitted model will be close to the likelihood of the null model. Hence the value of R^2_{McF} will be close to 0. On the contrary, if our model explains all of the variation in the outcome, it will have strong predictive capabilities. The probabilities of forecasting the correct outcome will be close to 1, and therefore the likelihood of the fitted model will also be close to 1. As a consequence, when taking logarithms we will obtain values close to 0, and therefore the value of R^2_{McF} will be close to 1. So we can conclude that the higher the value of the R^2_{McF} index, the more accurate our model will be.

On a different line, there are also tests that inform us of the contribution of the predictors to the model, i.e. the importance of these predictors. In this project we will focus on the Wald test Hastie et al. (2001). The Wald statistic is used to asses the significance of the coefficients of our model. This statistic is given by

$$W_j = \frac{\hat{\beta}_j}{\text{SE}^2_{\hat{\beta}_j}}, \tag{3.12}$$

where $\hat{\beta}_j$ is the regression coefficient and SE is its standard error. Asymptotically, this statistic is distributed as a χ^2 distribution and it the analogous to the t-test in linear regression. It is also important when dealing with estimation of the coefficients, to calculate confidence intervals to get an idea of how accurate our estimates are.

3.2.5 Additional remarks

The assumption of linearity may seem simple, but this model shows accurate results when applied to real-life problems in general. In addition, we can always use generalised additive models to allow non-linear dependence of the predictors, via splines Hastie et al. (2001). However, this method requires more computation than the linear method and in some cases can lead to overfitting, so we only focus on studying this linear model.

Once the setup is done and the model is presented, we may think of logistic regression as an analogous to linear regression for classification problems. However, there exist key differences between these 2 models. We highlight:

1. For logistic regression the conditional distribution $Y \mid X$ is a Bernoulli distribution whereas for the linear regression $Y \mid X$ is a Gaussian distribution;
2. For logistic regression predicts the probability of particular outcomes rather than the outcomes themselves. Therefore the logistic regression result is in the interval (0,1), whereas for linear regression we do not have any constraints on the result.

Lastly, it should be noted that one of the assumptions of this model is the independence of the predictors. Indeed, correlations between predictors may limit the applicability of this model or weaken its stability. Moreover as correlations increases, the estimators remain unbiased, but their variances increase, worsening the performance of the logistic regression method. In our case we can see in Figure 2.7 that our predictors are correlated. We will comment further on this fact during the model performance analysis.

3.3 Support vector machine

In this section we will explain how support vector machines (SVM) work. The general idea is again that we want to classify given data in two sets, a set with benign lymph nodes and a set with malignant lymph nodes. We divide the lymph nodes in one of the two sets based on the different predictor variables discussed in Section 2.3. Firstly, we will explain more about linear SVM classification in Subsection 3.3.1. Secondly, we will explain nonlinear SVM classification in Subsection 3.3.2. Finally, an explanation on the hyperparameters to be tuned is given in Subsection 3.3.3. The information used in these upcoming subsections has been obtained from Geron (2017) and Hastie et al. (2001). The results obtained with this model are presented in Section 4.3.

3.3.1 Linear SVM classification

The basic idea of Linear SVM is to find a hyperplane that separates the data in two sets A and B. This hyperplane is given by

$$\mathbf{w}^T\mathbf{x} + b = 0, \tag{3.13}$$

where $\mathbf{x}$ is the vector with the different variables. In our case, we have information on the primary tumour and the lymph node of a patient. If we can find such a hyperplane we call the data set linearly separable, i.e. for all data points in set A we have $\mathbf{w}^T\mathbf{x} + b \leq 0$ and for all data points in set B we have $\mathbf{w}^T\mathbf{x} + b \geq 0$. Our goal is to find an optimal hyperplane. A separating hyperplane is called optimal if the distance between the data points of both sets to the hyperplane is maximised, as shown by Figure 3.2 [2]. Automatically, the distance from the hyperplane to the closest point of set A is equal to the distance from the hyperplane to the closest point of set B.

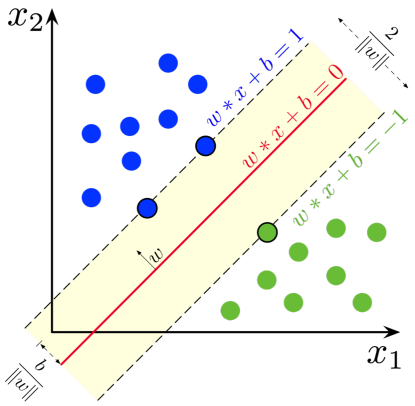


Figure 3.2: Visualisation of linear SVM

To find this optimal hyperplane we write this as an optimisation problem and use our training data to solve it. In our case we want to predict whether a lymph node is benign or malignant. For the training data, we know if a certain lymph node is benign or malignant. Benign and malignant are our sets A and B. If a lymph node is benign we want that the inequality $\mathbf{w}^T\mathbf{x} + b \leq -1$ holds. And if a lymph node is malignant we want that the inequality $\mathbf{w}^T\mathbf{x} + b \geq 1$ holds.

[2] https://en.wikipedia.org/wiki/Support-vector_machine [Last Accessed 11-05-2021]

Notice, the right hand side of the inequalities was zero in the previous paragraph and now it is one. This is, because we want to find two parallel hyperplanes. With the distance in between the hyperplanes as large as possible. In the middle of these two hyperplanes there will be the optimal hyperplane, see Figure 3.2.

To write these two constraints as one constraint that holds for all data points i in the training data we introduce a variable $k^{(i)}$, with $k^{(i)} = -1$ if a lymph node is benign and $k^{(i)} = 1$ if a lymph node is malignant. This results in the constraint:

$$k^{(i)}(\mathbf{w}^T\mathbf{x}^{(\mathbf{i})} + b) \geq 1, \forall \text{ data points } i \text{ in the training data.} \tag{3.14}$$

As mentioned before we want to maximise the distance between the two hyperplanes ($\mathbf{w}^T\mathbf{x} + b = 1$ and $\mathbf{w}^T\mathbf{x} + b = -1$). We will first calculate the distance between these two hyperplanes [3]. Notice that these two hyperplanes are parallel. The distance between these two hyperplanes is exactly the distance between v on the hyperplane $\mathbf{w}^T\mathbf{x} + b = 1$ and the hyperplane $\mathbf{w}^T\mathbf{x} + b = -1$. Because v is on the hyperplane we have $\mathbf{w}^T\mathbf{v} + b = 1$. We also know that $\frac{\mathbf{w}}{||\mathbf{w}||}$ is a unit normal vector of both hyperplanes. Combined this gives that there exists a $s \in \mathbb{R}$ such that

$$\mathbf{w}^T\left(\mathbf{v} + s\frac{\mathbf{w}}{||\mathbf{w}||}\right) + b = -1. \tag{3.15}$$

We can rewrite this as

$$\begin{aligned}
\mathbf{w}^T\mathbf{v} + s\frac{\mathbf{w}^T\mathbf{w}}{||\mathbf{w}||} + b &= -1 \\
s\frac{||\mathbf{w}||^2}{||\mathbf{w}||} + 1 &= -(\mathbf{w}^T\mathbf{v} + b) \\
s||\mathbf{w}|| + 1 &= -1 \\
s||\mathbf{w}|| &= -2 \\
s &= -\frac{2}{||\mathbf{w}||}.
\end{aligned} \tag{3.16}$$

From this we can conclude that the distance between the two hyperplanes is $|s| = |-\frac{2}{||\mathbf{w}||}| = \frac{2}{||\mathbf{w}||}$.

We want to maximise the distance between the two hyperplanes. Which is the same as maximising $\frac{2}{||\mathbf{w}||}$. Which is equivalent to minimizing $||\mathbf{w}||$. To make the optimisation process easier we will minimise $\frac{1}{2}\mathbf{w}^T\mathbf{w}$. Combined with the constraints (3.14), we get the optimisation problem for linear SVM:

$$\begin{aligned}
&\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^T\mathbf{w} \\
&\text{subject to } k^{(i)}(\mathbf{w}^T\mathbf{x}^{(\mathbf{i})} + b) \geq 1, \forall \text{ data points } i \text{ in the training data.}
\end{aligned} \tag{3.17}$$

[3] https://math.stackexchange.com/questions/1305925/why-is-the-svm-margin-equal-to-frac2-mathbfw [Last Accessed 16-05-2021]

We can extend this to a method where we do allow some data points to violate the two hyperplanes a little bit. When we add this to our model we need to add two more variables. Firstly, we introduce the slack-variable $\zeta^{(i)}$ which indicates how much data points i violate the constraint (3.14). Secondly, we introduce C. This variable indicates if we want to focus more on minimizing the total violation of the constraints (C has a high-value) or on maximizing the distance between the two hyperplanes (C has a low-value). This gives the primal optimisation problem

$$\begin{aligned} &\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m}\zeta^{(i)} \\ &\text{subject to } k^{(i)}(\mathbf{w}^T\mathbf{x^{(i)}} + b) \geq 1 - \zeta^{(i)}, \forall \text{ data points } i \text{ in the training data,} \\ &\qquad \zeta^{(i)} \geq 0, \forall \text{ data points } i \text{ in the training data} \end{aligned} \tag{3.18}$$

with m the number of data points in the training data. To solve this optimisation problem, it is more convenient to rewrite equation (3.18) to its dual form. The primal problem is quadratic with linear inequality constraints, and therefore it is a convex optimisation problem. Furthermore, since the inequality constraints are continuously differentiable and convex, both the dual and primal problem have the same solution. The dual form is

$$\begin{aligned} &\max_{\alpha} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha^{(i)}\alpha^{(j)}k^{(i)}k^{(j)}\mathbf{x}^{(i)T}\mathbf{x^{(j)}} - \sum_{i=1}^{m}\alpha^{(i)} \\ &\text{subject to } \alpha^{(i)} \geq 0, \forall \text{ data points } i \text{ in the training data and } \sum_{i=1}^{m}\alpha^{(i)}k^{(i)} = 0. \end{aligned} \tag{3.19}$$

The solution of the dual can be expressed with the equations

$$\hat{\mathbf{w}} = \sum_{i=1}^{m}\hat{\alpha}^{(i)}k^{(i)}\mathbf{x}^{(i)} \tag{3.20}$$

$$\hat{b} = \frac{1}{n_s}\sum_{\substack{i=1 \\ \hat{\alpha}^{(i)}\geq 0}}^{m}(k^{(i)} - \hat{\mathbf{w}}^T\mathbf{x}^{(i)}), \tag{3.21}$$

where the coefficients $\alpha^{(i)}$ are nonzero only for those training points i for which the constraints of equation (3.18) are exactly met. These training points are called the support vectors. Given the solutions for $\hat{\mathbf{w}}$ and $\hat{\beta}$ the final decision function can be written as

$$\hat{G} = \text{sign}(\hat{f}(x)) \tag{3.22}$$

$$= \text{sign}(\hat{\mathbf{w}}^T x + \hat{b}). \tag{3.23}$$

3.3.2 Nonlinear SVM Classification

In Subsection 3.3.1 we looked at Linear SVM Classification. In many cases the linear classifiers works well and are efficient. However, many datasets are not linearly separable. In that case the dataset can

be transformed by adding more features such that the transformed dataset can be linearly separated. Intuitively this goes as follows. Once a suitable transformation function $h_m(x), m = 1, \ldots, M$ is selected, the procedure is the same as in the linear case. So, a (linear) SVM classifier will be fitted using the input features $h(x_i) = (h_1(x_i), h_2(x_i), \ldots, h_M(x_i)),\ i = 1, \ldots, N$ and produces the (nonlinear) hyperplane function $\hat{f}(x) = \hat{\mathbf{w}}h(x)^T + \hat{b}_0$. The classifier is $\hat{G}(x) = \text{sign}(\hat{f}(x))$. This can be visualised as in Figure 3.3 obtained from Rawat (2020). This intuitive idea can work well for lower dimension spaces. However, when we are dealing with higher dimensional spaces the computational complexity increases and the computations become more expensive. To avoid this, a trick, called the Kernel trick, can be applied.

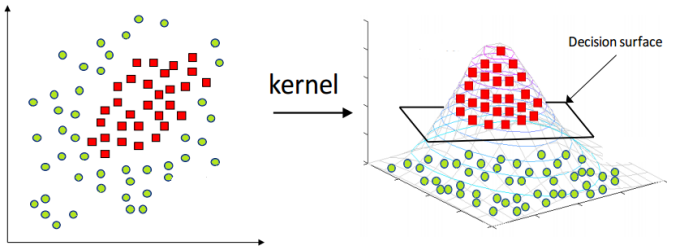


Figure 3.3: Transforming data-points to a higher dimensional space such that a linear SVM classifier can be applied

An easy way to explain this kernel trick is by looking at an example. So suppose you want to transform a two-dimensional training set to a three dimensional feature space and subsequently train a linear SVM classifier to the transformed training set. The mapping can be done by

$$\phi(\mathbf{x}) = \phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{bmatrix}. \tag{3.24}$$

Now if we transform two 2D vectors $\mathbf{a}$ and $\mathbf{b}$ with the function above, the dot product of these transformed vectors can be computed as

$$\begin{aligned} \phi(\mathbf{a})^T\phi(\mathbf{b}) &= \begin{bmatrix} a_1^2 & \sqrt{2}a_1a_2 & a_2^2 \end{bmatrix} \begin{bmatrix} b_1^2 \\ \sqrt{2}b_1b_2 \\ b_2^2 \end{bmatrix} = a_1^2b_1^2 + 2a_1b_1a_2b_2 + a_2^2b_2^2 \\ &= (a_1b_1 + a_2b_2)^2 = \left(\begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}\right)^2 = (\mathbf{a}^T\mathbf{b})^2. \end{aligned}$$

From this we can see that we actually don't need to first transform the vectors and subsequently calculate the dot product of the transformed vectors. We only have to compute the function $K(\mathbf{a},\mathbf{b}) = (\mathbf{a}^T\mathbf{b})^2$, which is called the kernel.

If we reconsider the optimisation problem of previous section, and substitute the transformed feature

vectors, the dual objective function has the form

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha^{(i)}\alpha^{(j)}k^{(i)}k^{(j)}\phi(\mathbf{x}^{(i)T})\phi(\mathbf{x}^{(j)}) - \sum_{i=1}^{m}\alpha^{(i)}.$$

In this function we can replace $\phi(\mathbf{x}^{(i)T})\phi(\mathbf{x}^{(j)})$ by the kernel function $K(\mathbf{a},\mathbf{b})$, for particular choices of $\phi()$. Popular choices of kernel function are the polynomial kernel and the Gaussian RBF kernel. The polynomial kernel adds polynomial features. The Gaussian RBF kernel adds feature by using a similarity function. The kernel functions are:

$$\begin{aligned} \text{Polynomial}: &\quad K(\mathbf{a},\mathbf{b}) = (\gamma\mathbf{a}^T\mathbf{b} + r)^d \\ \text{Gaussian RBF}: &\quad K(\mathbf{a},\mathbf{b}) = \exp\{-\gamma||\mathbf{a}-\mathbf{b}||^2\}. \end{aligned}$$

In further analysis of the SVM we will focus on the linear SVM and these kernelized SVMs. To get more insight on these SVM classifiers on our dataset we created Figure 3.4. Here you see the classifiers fitted for only two features, namely the primary tumour SUV and the lymph node SUV. Hence, Figure 3.4 illustrates the "working principle" of the different classifiers.

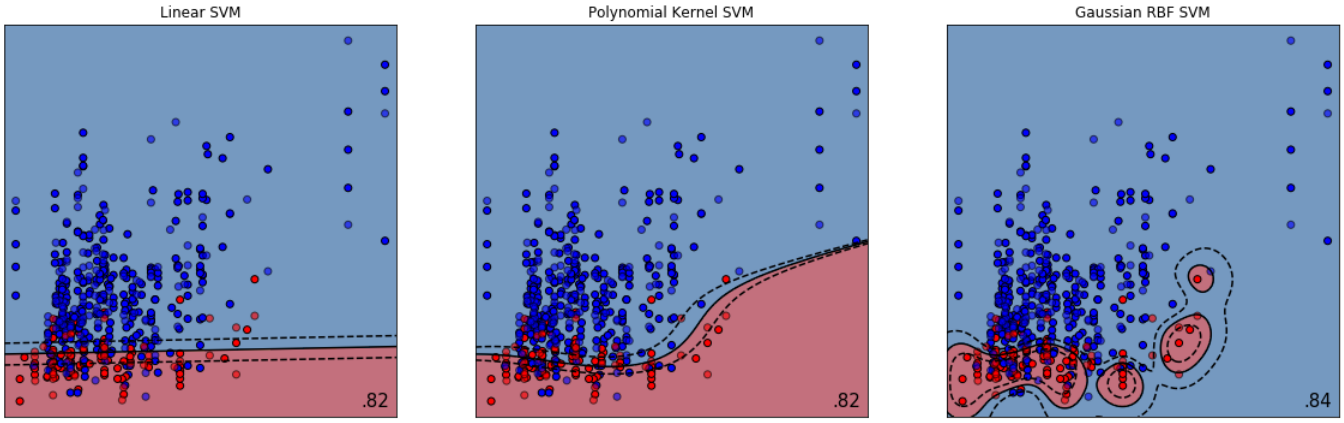


Figure 3.4: Different SVM classifiers fitted on two features, primary tumor SUV and the lymph node SUV

3.3.3 Hyper parameter explanation

There are a couple of hyperparameters which can be tuned when we fit an SVM classifier on the training data. Both the linear as the polynomial as the Gaussian SVM classifiers contain the hyperparameter C. As said in the first paragraph, the parameter C indicates if we want to focus more on minimising the total violation of constraints or maximising the distance between the two hyperplanes. In other words, a small value of C, for instance $C = 0.1$ or $C = 1$, gives a larger distance between dashed lines of Figure 3.4 and a larger value, such as $C = 100$ results in a smaller distance between the dashed lines. In addition, the polynomial kernel SVM has the hyperparameters γ, the coefficient r and the degree d. The degree parameters obviously indicates which degree polynomial is used. If

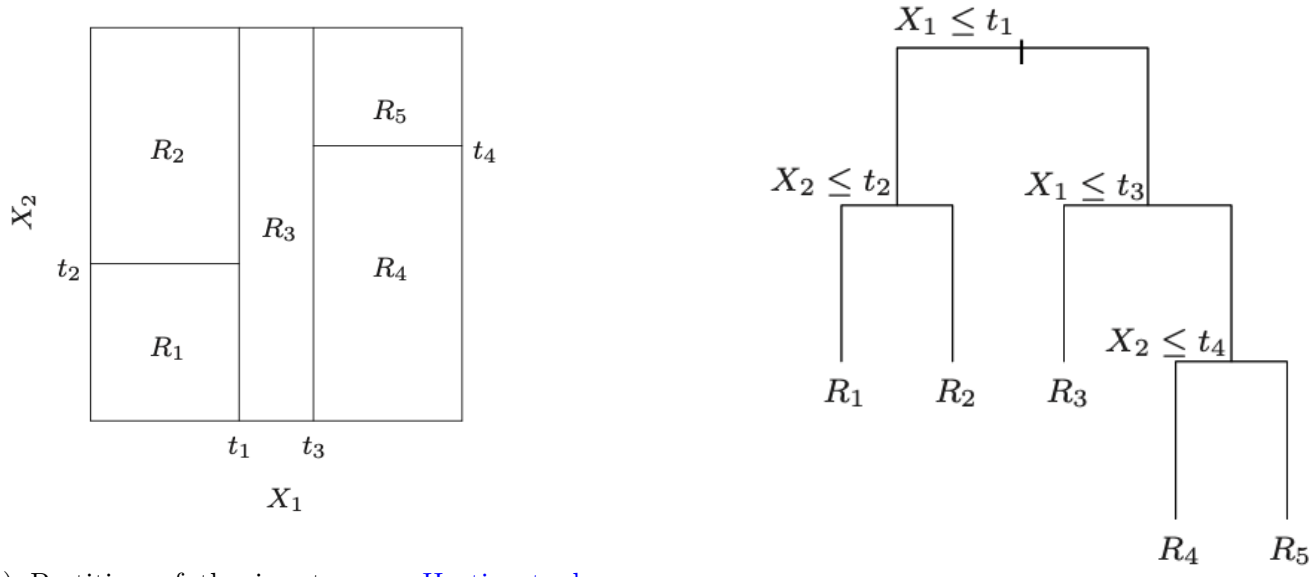
a model is underfitting the training data, we can try to increase the degree. However, using a degree too high, can result in overfitting of the model. The coefficient r regulates how much the model is influenced by high-degree polynomials. The parameter γ which is also used as hyperparameter for a Gaussian SVM classifier, controls the curvature of the decision boundary. For a large value of γ, the curves of the decision boundary are narrower resulting in more wiggles around individual data points. If the value is too high, the SVM classifier likely will be overfitting. Reducing the γ-value smoothens the decision boundary.

3.4 Decision tree methods

In this section, the theory behind decision tree methods are discussed. First, Subsection 3.4.1 discusses classification trees in general. Next, Subsection 3.4.2 and Subsection 3.4.3 discuss the theory behind the ensemble methods of bagging and random forests, respectively. The information contained in this subsection is obtained from Hastie et al. (2001). The results obtained with this model are presented in Section 4.4.

3.4.1 Decision trees

Basically, when building a decision tree, the dataset is split into homogeneous subgroups. Then, the decision trees fits a constant, i.e the mean of the responses within the subgroup, for each subgroup. These subgroups are made through simple questions with binary answers such as, is the PV greater than 700 (yes or no?). A visualisation of this is given in Figure 3.5.



(a) Partition of the input space Hastie et al. (2001)

(b) Tree representation Hastie et al. (2001)

Figure 3.5: Example of the binary partitioning of trees

This partitioning is done until a certain stopping criteria is reached. Let's say the input space is partitioned into M regions $R_1, R_2, \ldots, R_M$. Then the prediction model becomes

$$\hat{y}(x) = \sum_{m=1}^{M} \hat{y}_m 1_{\{x \in R_m\}}. \tag{3.25}$$

For classification trees, the proportion of data points from class k in node m corresponding to the

correct region R_m with N_m points can be computed by

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{i:x_i \in R_m} 1_{y_i=k}. \tag{3.26}$$

Each splitting minimises that sum of losses for all possible splits. For this, measures are needed which are suitable for classifying problems. Useful measures could be: misclassification error, entropy/deviance, Gini index. When using the package `rpart` in r [4], this splitting is done automatically.

There are two methods which can improve the practical performance of the tree, namely 'pruning' the original tree and 'ensemble methods'.

Pruning is a data compression technique that reduces the size of decision trees by removing sections or tree 'parts' that are non-critical and redundant to classify instances. Pruning a tree results in a smaller variance and improves the predictive accuracy. The optimal pruning value can be read from a pruning complexity plot. This shows the cross-validation error for various tree sizes, and gives an idea on how big the tree size should be. In the results, both a full grown tree and a pruned tree can be found. The pruned tree is eventually used to get our results and conclusions. Figure 3.6 [5] shows an example of a unpruned and pruned tree.

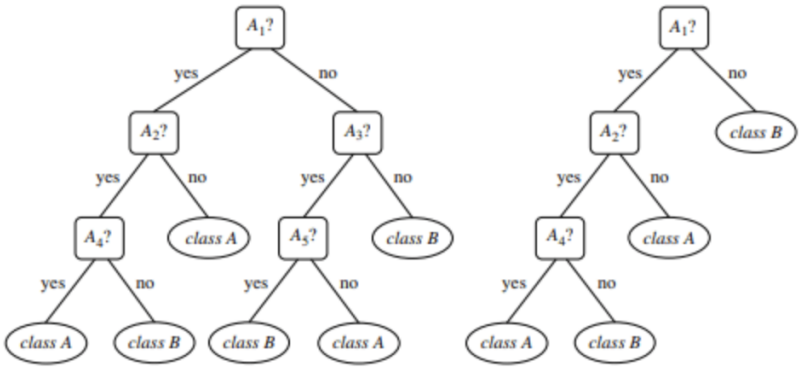


Figure 3.6: On the left an unpruned tree and on the right a pruned tree

Ensemble methods combine multiple decision trees to produce better predicting performance compared to using a single decision tree. Bagging is a ensemble method which is explained in Subsection 3.4.2. Random forest is an extension of bagging, in Subsection 3.4.3 it is explained.

3.4.2 Bagging

Bootstrap aggregating or 'bagging' in short, combines multiple prediction models. Bagging averages the predictions of all the multiple models, this results in a reduced variance compared to a single

[4] https://github.com/bethatkinson/rpart [Last Accessed 14-05-2021]

[5] https://www.softwaretestinghelp.com/decision-tree-algorithm-examples-data-mining/ [Last Accessed 26-05-2021]

decision tree. When using a train and test set with bagging, the method produces multiple new training sets by sampling from the original train set. From all those new training sets the average is taken and compared to the test set.

When generating new training sets, the size can be determined. If the size of the produced new training sets are the same as the size of the original training set, it can be expected that ca. 60% of the generated sets are unique samples of the original training set. The other ca. 40% are duplicates of the generated new training sample sets. Bagging is a good model for unstable methods, such as classification trees, because it reduces the variance. It is common to use for estimation 50-500 trees. We start with using 200 trees for the bagging model. This means that 200 bagged unpruned trees are build, in order to keep the bias low and reduce the variance. The function `randomForest` [6] is used for bagging. It calculates and uses the optimal size for the generated training trees.

3.4.3 Random forest

Random forest is an extension of the bagging method. The random forest method also averages multiple decision trees, but it also adds some random noise to the data which results in less correlation between the trees. When bagging a model, there can be some correlation between the trees because the grown trees are very similar. The random forest model reduces this. Specifically, the splitting moments only depend on the variance within a random subset, instead of the whole set. This parameter which determines how big this random group will be, is called the tuning parameter.

Random forest performs split-variable randomisation where each time a split is performed, the split variable is limited to a random subset m_{try} of the p original features. In our case $p = 5$. For regression problems, the preferred value of m_{try} is floor $\left(\frac{\text{number of features}}{3}\right)$.

The algorithm for random forest starts with fitting a random forest to the data. For each data point, the OOB (out-of-bag) error is calculated and averaged. Then, the variable importance score can be determined by averaging the difference in OOB errors before and after the permutations and normalising it by using the standard deviation of these differences. Higher scores mean that the features have more importance. In Subsection 4.4.1, the variable importance of all decision tree models are discussed.

[6] https://cran.r-project.org/web/packages/randomForest/index.html [Last Accessed 30-05-2021]

3.5 Measures for goodness-of-fit

For comparing the results, it is important to use appropriate performance measures for goodness-of-fit. For some specific models, specific measures can be used which are useful for that specific model. However, as we are using four different models, it is important to have measures which are useful for all three of them. Therefore, we use two different measures:

- Confusion matrix Stehman (1997);
- AUC-ROC Curve [7].

In this section, we will elaborate on both measures and discuss why these two are appropriate for our research. The confusion matrix measure is discussed in Subsection 3.5.1 and the AUC-ROC curve is discussed in Subsection 3.5.2.

3.5.1 Confusion matrix

For classification problems, it is common to use the confusion matrix in order to measure how effective the model is. Basically, it compares the predicted value and the validation set. It's output is a table which combines the predicted values and the actual values and classifies them into four categories:

- True Positive (TP);
- True Negative (TN);
- False Positive (FP);
- False Negative (FN).

In this report, it is also visualised as a table. This table looks like Figure 3.7 [8].

This matrix is very useful for specificity, sensibility and accuracy. It also can be interpreted by the preference of a person. For our research for example, it might harm less to have a false positive instead of a false negative. Since these numbers are also provided within this measure, it is a very useful tool for our problem. From this matrix, the accuracy of the model can be calculated. Also other measures such as sensitivity (or true positive rate) and specificity can be found. For all measures, it is important that the percentage is as high as possible. These measures can be calculated as follows.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \tag{3.27}$$

[7] https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 [Last Accessed 25-05-2021]
[8] https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

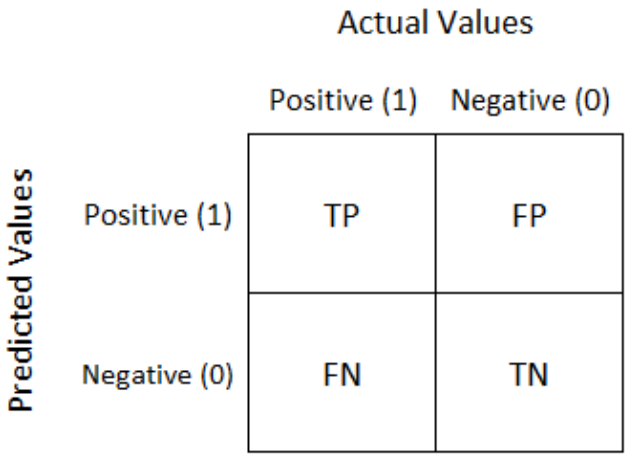


Figure 3.7: Visualisation of the confusion matrix

which measures out of all positive classes, how much was predicted correctly.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}}, \tag{3.28}$$

which measures out of all negative classes, which were correctly predicted negative.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{Total}}, \tag{3.29}$$

which measures out of all classes, how many were predicted correctly.

Sensitivity is considered to be one of the most important parameters in this analysis. The reason for this is the fact that specifically for this case it is more important to accurately classify malignant lymph nodes than benign lymph nodes. A low sensitivity indicates a large number of false negatives compared to the number of true positive predictions. Rejecting a true positive is unacceptable for this case and can be dangerous in the setting of medical prediction, since performing a biopsy on a patient whose lymph nodes are actually benign is less harmful than not performing a biopsy on a patient whose lymph nodes are actually malignant. As a result, lower specificity is of less concern and its mainly important that the models strive for high sensitivity.

Furthermore, though less extreme than sensitivity, a high specificity is also important. The downside to a low specificity is that such models may result in a large number of patients going into treatment unnecessarily, since low specificity implies a large number of false positives. This may result in unnecessary costs and time spent on a patient that may not need the treatment. Furthermore, unnecessary risks may even be introduced on the patient by performing a biopsy that is not needed, though these risks are extremely limited [9]. To conclude, we are not as concerned with a low specificity as a low sensitivity, but an extremely low specificity is still undesirable.

From these three features, it can be hard to compare two models with for example low specificity and high sensitivity (or the other way around). Therefore, an F-score can be calculated in order to

[9] https://www.healthline.com/health/biopsy [Last Accessed 25-05-2021]

make the models comparable. The F-score can be calculated by

$$\text{F-score} = \frac{2 \cdot \text{sensitivity} \cdot \text{specificity}}{\text{sensitivity}+\text{specificity}}. \tag{3.30}$$

This is the harmonic mean of the specificity and the sensitivity. The score punishes extreme values more, and is therefore useful for our problem. For significance testing, it can be used as well by applying the F-test. In our case, all models are significant and therefore, we won't use this score for comparing the performance of the models.

An advantage of the confusion matrix, is that it is easily interpretable. However, to optimise the goodness-of-fit measure, it is very useful to take the AUC/ROC curve into account. This combines the different features stated above, into one measure. More information on the ROC/AUC curve is given in Subsection 3.5.2.

3.5.2 AUC-ROC curve

For visualizing the performance of the classification models, the AUC (Area Under The Curve) - ROC (Receiver Operating Characteristics) curve can be used. It is the most common goodness-of-fit measure for classification problems. ROC is a curve that is based on probabilities and AUC shows the degree of separability, which means that is tells us how capable the model is in distinguishing classes. The higher the AUC, the better the model predicts true positives and true negatives, which in our research very relevant. The ROC curve plots the sensitivity versus the false positive rate (FPR). In Figure 3.8, this is graphically shown.

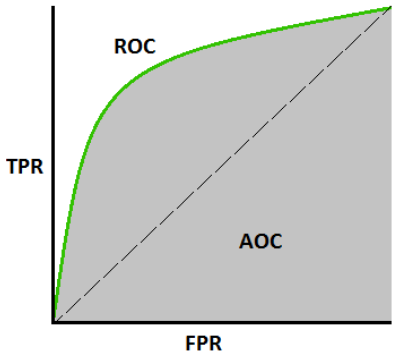
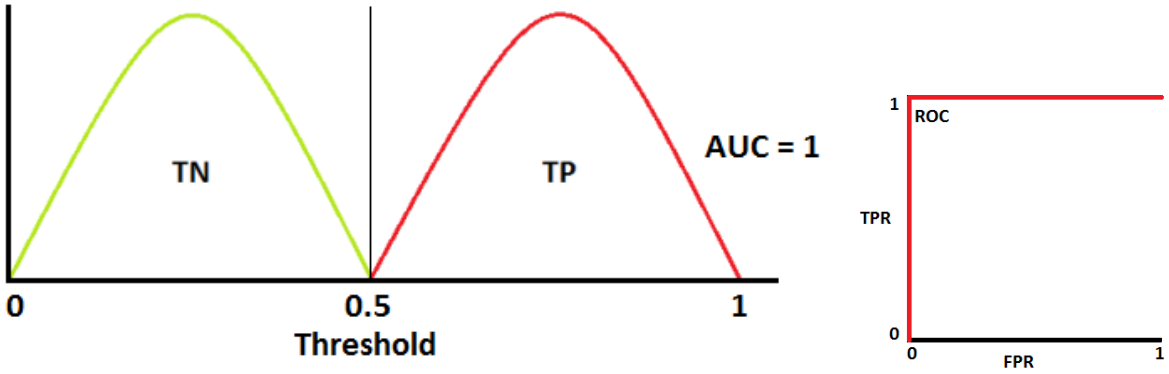


Figure 3.8: Visualization of the AUC-ROC curve

The sensitivity is calculated such as in the previous section. The false positive rate is found by

$$1 \text{ - Specificity} = \text{FPR} = \frac{\text{FP}}{\text{TN}+\text{FP}}.$$

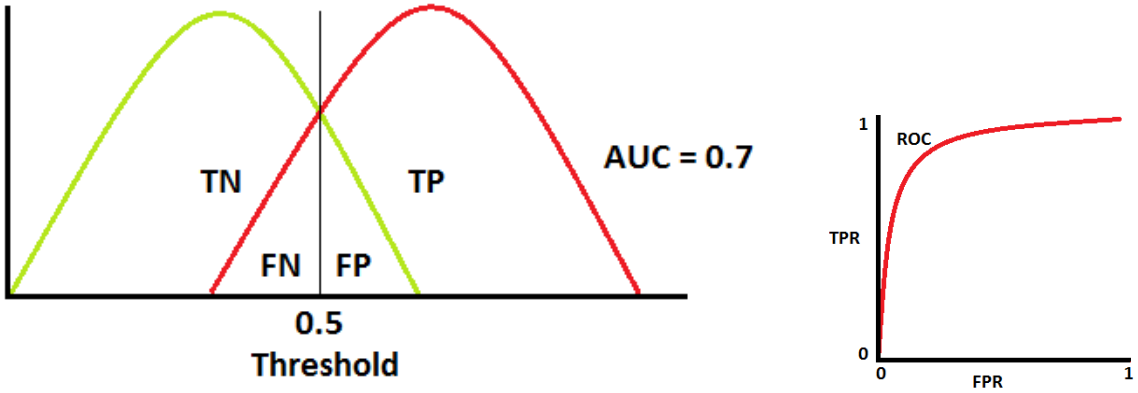
The closer the AUC is to one, the better the model is, since it distinguishes the classes well. When the AUC equals 0.5, it means the model does not correctly separate the different classes at all. To show how to interpret the ROC curve, Figure 3.9 can be used.



(a) (b)

Figure 3.9: Perfect situation for ROC curve

Here, Figure 3.9a and Figure 3.9b show the ideal situation with only true positives and only true negatives. As we see, the curves do not overlap which means the models separates the classes perfectly. More realistically, there can be false positives and false negative. The ROC curve shows this as in Figure 3.10.



(a) (b)

Figure 3.10: Realistic situation for ROC curve

As we can see, the two distributions overlap, here, type 1 and type 2 errors occur. When the AUC is for example 0.8, it means that there is a probability of 80% that the model distinguishes the positive and negative classes correctly. When the two curves totally overlap, the ROC curve looks like the $x = y$ line, which is the worst situation. In that case, the AUC has a value of 0.5 and the model has no capacity of distinguishing positive and negative classes.

Since for our models, we only have two classes in which the predictions are classified (namely positive and negative), one plotted curve will give us enough information. For classifying problems, we can make the model more sensitive which will give us more positive values. This has as a consequence that the model becomes less specific, since it will also give more false positives. When false positives are desired above false negative, it can be an option to lower the threshold.

In conclusion, for easy interpretation, the confusion matrix is useful. The ROC-AUC curve shows

similar features, however, it combines specificity and sensitivity. Both measures give a good indication on the performance of our measures. For our research, we will however use the confusion matrix and elaborate on the sensitivity and specificity score because it gives the reader the space to interpret the results as well. One can for example value a more sensitive over a specific model. This can lead to more false positives / negatives. We also provide the AUC score but for comparison of the models, the accuracy and sensitivity are hegemonic.

4 Model performance analysis

In this chapter, the performance of each model described in Chapter 3 is analysed in terms of the measures listed in Section 3.5. For each model, the performance analysis consists of two parts: a variable analysis, and a performance analysis. The models are presented in the same order as in Chapter 3. First, the performance of the thresholding model is analysed in Section 4.1. Next, the logistic regression model is evaluated in Section 4.2. Following this, the performance of the SVM models is evaluated in Section 4.3, after which the performance of the decision tree methods is assessed in Section 4.4. An analysis on the different hospitals is performed in Section 4.5. Finally, as mentioned in Subsection 2.4.6, all models are evaluated using the exact same train and test data. In order to assess the extent of change in accuracy with a change in train test data split, a robustness analysis is carried out for all models in Section 4.6.

4.1 Thresholding

In this section the performance of the thresholding model described in Section 3.1 is analysed. This model only uses a single variable, so it does not make sense to consider the variable analysis. However, as we will see later in Section 4.2 - Section 4.4, the LSUV variable is the most important variable for the these models. For this reason, the thresholding is based on LSUV.

The first step is to set the threshold level to obtain a prediction criterion for our test set. We choose this level so that the accuracy of our train set is the maximum possible. For our train set we obtain:

$$\boxed{\text{Threshold}_{\text{level}} = 3.3426}$$

which give us Accuracy $= 0.8465909$. However, the results we are interested in are those of the test set, independent of the cut-off value we have chosen. Evaluated in our test data set, we obtain:

$$\boxed{\text{Accuracy} = 0.8486842}$$

Despite the simplicity of this model, the results obtained seem accurate. In order to corroborate this last statement, a study on robustness analysis is conducted in Section 4.6.

Based on the $\text{Threshold}_{\text{level}}$ the corresponding confusion matrix for the Thresholding model is given by Table 4.1

Table 4.1: Confusion Matrix for Thresholding model.

		True	
		Malignant	Benign
Predicted	Malignant	100	6
	Benign	17	29

Once the confusion matrix is computed, we can calculate the sensitivity and the specificity as given by (3.27) and (3.28), respectively. The values that we obtain for this criterion are:

$$\boxed{\text{Sensitivity} = 0.8547009}$$

$$\boxed{\text{Specificity} = 0.8285714}$$

It can be observed that the sensitivity is higher than the specificity for the thresholding model, which is desirable.

Lastly, by varying the value of $\text{Threshold}_{\text{level}}$, we can compute the different values taken by the sensitivity and the specifity. Plotting these values in Figure 4.1 we obtain the ROC curve of the Thresholding model.

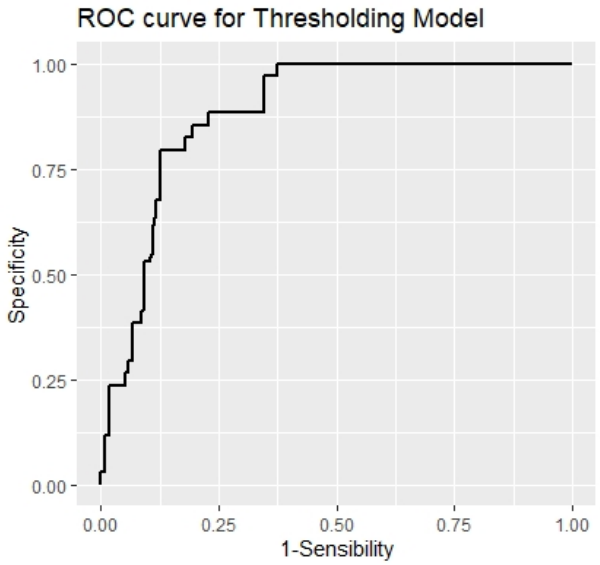


Figure 4.1: ROC curve for the Thresholding model.

The area under this previous curve is:

$$\boxed{\text{AUC threshold} = 0.8957265}$$

4.2 Logistic regression

In this section, the performance of the logistic regression model described in Section 3.2 is analysed. First, the variables and their respective importances in the model are evaluated in Subsection 4.2.1. Next, the actual performance of the model is assessed in Subsection 4.2.2. To carry out this analysis we have consulted Hosmer & Lemeshow (2000) and James et al. (2013).

4.2.1 Variable analysis

After having developed the logistic regression method in Section 3.2 from a theoretical point of view, we move on to a practical approach. In this section we apply the logistic regression method to the lymph node classification problem. The results shown below have been obtained using R. For R, we have mainly used the library `stats` [1].

First of all we, divide our data in training and test data set, using the stratified splitting described in Section 2.4. We build the model using the training test, but we will use the test data set to asses its performance. Using a generalised linear model to fit our logistic regression model, we obtain:

```
Call:
glm(formula = as.numeric(Pathological.Result) ~ Primary.Tumour.Volume +
    Primary.Tumour.Mean.SUV.Value + Lymph.Node.Volume + Lymph.Node.Mean.SUV.Value +
    Lymph.Node.Mean.HU.Value, family = "binomial", data = train_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4211   0.0006   0.1302   0.5238   2.0180

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -6.128e+00  1.563e+00  -3.921 8.82e-05 ***
PV                          -4.152e-06  2.043e-06  -2.032  0.04217 *
PSUV                         2.005e-02  8.539e-02   0.235  0.81431
LV                           2.622e-04  8.654e-05   3.030  0.00244 **
LSUV                         1.375e+00  2.138e-01   6.432 1.26e-10 ***
LHU                          1.623e-03  1.407e-03   1.154  0.24857
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 382.15  on 351  degrees of freedom
Residual deviance: 226.18  on 346  degrees of freedom
```

[1] https://www.rdocumentation.org/packages/stats/versions/3.6.2 [Last Accessed 10-05-2021]

```
AIC: 238.18

Number of Fisher Scoring iterations: 8
```

First, the formula indicates the predictors that we are taking into account to feed the model. These predictors are the ones described in Section 2.3. Second, we see a detailed description of the estimators of the predictors, with its standard deviations and the p-value for the Wald test. We highlight the following facts:

- As mentioned in Section 3.2, the intercept describes the null-model. Therefore, the chances of a lymph node being malignant, when the predictors are null (among them LSUV and LV) is

$$\mathbb{P}(Y = 1 \mid \text{null model}) = e^{-6.128} \approx 0.00218.$$

- The magnitude of the estimator must be interpreted taking into account the magnitude of the associated predictor. This can be illustrated as follows: The LV value range is $[147, 97120]$ and the LSUV value range is $[1.616, 12.940]$. Varying one unit in the first case is near negligible, while varying one unit in the second case means varying by almost 10% of the range, which can have great effects.

- Actually, the informative measure of the significance of the predictors is found in the p-value of the Wald test, in the last column. The lower the p-value, the more important role a predictor plays in our model. In this case, the most important variables are: LSUV, LV and the intercept. It is curious to see that the PV measure associated with tumours have little influence on their corresponding lymph nodes. Moreover, the p-value of LHU and PSUV is so large that these variables will have practically no effect on the construction of our model.

- Finally, it is also interesting to analyse the sign of the estimators. As mentioned in Section 3.2, if the regressor coefficient is positive, as the associated predictor increases, the probability that the lymph node is malignant increases. This is what happens with LV and LSUV, as expected. On the other hand, if the regressor coefficient is negative the opposite is true. In this analysis, we were struck by the fact that the remaining predictors are inversely proportional, contrary to what logic would tell us. It is unclear whether this is due to an instability of our model due to possible correlations between our predictors. Or if it is a medical-biological fact that escapes our knowledge as mathematicians.

- We might think that one way to improve the model would be to eliminate uninformative variables. However, due to the already small number of predictors, eliminating some of them could lead to a deterioration of our model.

The next step is to calculate the confidence intervals for these regressors. Using profiled log-likelihood we obtain that the regressors in (3.2) are:

```
                                2.5 %        97.5 %
(Intercept)                -9.337953e+00 -3.198136e+00
PV                         -8.367986e-06 -2.525525e-07
PSUV                       -1.456793e-01  1.911211e-01
LV                          1.101714e-04  4.485110e-04
LSUV                        9.840288e-01  1.825330e+00
LHU                        -1.147431e-03  4.368017e-03
```

Exponentiating these estimators as we did in (3.6) we obtain the odd-ratios, which is an easier measure to interpret. Including the previous confidence intervals we obtain:

```
                                 OR        2.5 %     97.5 %
(Intercept)                0.002180483 0.0000880194 0.04083826
PV                         0.999995848 0.9999916320 0.99999975
PSUV                       1.020257424 0.8644348430 1.21060599
LV                         1.000262272 1.0001101775 1.00044861
LSUV                       3.956245913 2.6752125019 6.20484445
LHU                        1.001624727 0.9988532267 1.00437757
```

As mentioned above, these results must be interpreted taking into account the magnitude of the associated predictor.

Once we have obtained the model, we are ready to apply it to our test data to see how accurate the predictions are. Applying the model, we obtain a series of probabilities for each of the observations made.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.07369 0.58326 0.92000 0.77280 0.99754 1.00000
```

To understand the influence of the predictors on the prediction probabilities, it is convenient to perform a series of plots. First we choose the predictor that we are interested in. Then we create a new data set, where:

- We vary the chosen predictor between the min value and max value that we had in our training set.
- We set the rest of the predictor constants. In our case we have chosen these values as the mean of the training data values we had.

Lastly, we apply our logistic regression model, to compute the predicted probabilities and we plot them against the range values of the chosen predictor. The predictors that we analyse are: Lymph node mean SUV value, Lymph node volume and Primary Tumour SUV Value. Moreover we include the confidence intervals of the predicted probabilities. For the Lymph node mean SUV value, Figure 4.2 is obtained.

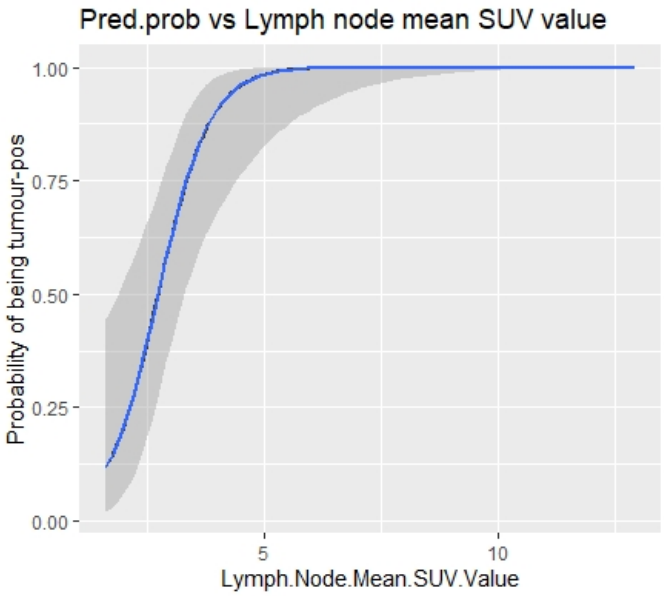


Figure 4.2: Influence of the Lymph node mean SUV value in the predicted probabilities.

As expected, the lymph node mean SUV value is the most important variable in our model, so by varying it, the predictions we obtain will vary greatly. For the Lymph node volume, Figure 4.3 is obtained.

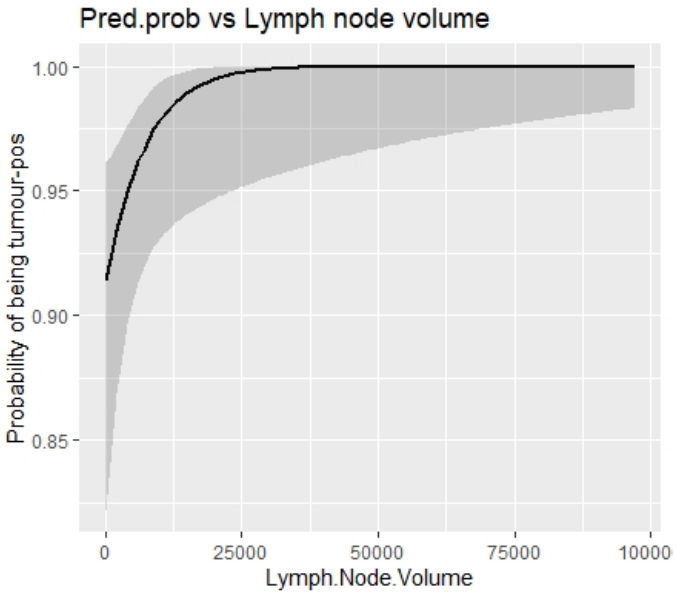


Figure 4.3: Influence of the Lymph node volume in the predicted probabilities.

In this case the graph is largely affected by the lymph node mean SUV value. As mentioned above we take this value constant and equal to the mean value obtained in the training set. In the training set there are many more positive than negative pathological results. Then the mean of the LSUV predictor, will be relatively high, and it will lead us to high probabilities. Despite this fact, we see that as we decrease the value of the lymph node volume, the probabilities decrease. For the Primary Tumour SUV Value, Figure 4.4 is obtained.

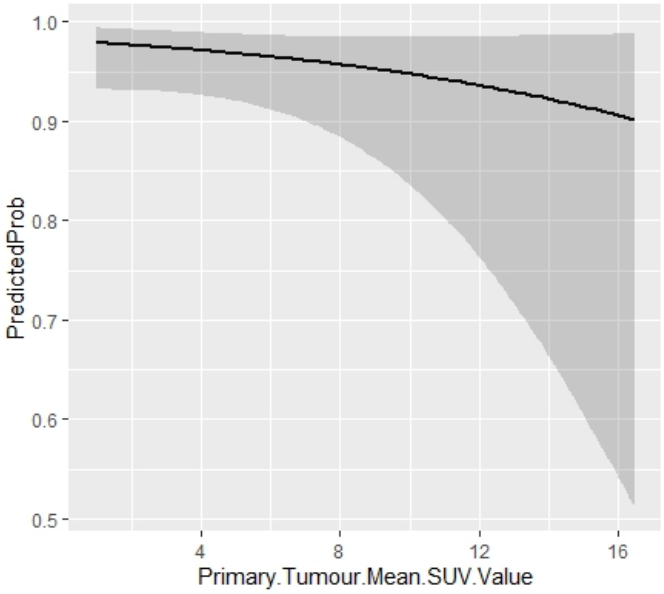


Figure 4.4: Influence of the tumour mean SUV value in the predicted probabilities.

With respect to Figure 4.4, it is worth noting the inversely proportional relationship that we obtained for the estimator of this predictor. Moreover, due to the low importance of this predictor in our model, the results are poor and the confidence intervals are very large, especially for high values of the predictor.

4.2.2 Performance analysis

The next step is to choose the criteria to be used for binary prediction purposes.

- Using $\alpha = 0.5$ in the criterion described in (3.7), we obtain: $\boxed{\text{Accuracy} = 0.8552632}$.
- Using $\alpha = 0.65$ the best results are obtained. In this case we have: $\boxed{\text{Accuracy} = 0.8815789}$.
- Using random samples for a Bernoulli distribution and setting a seed (set.seed(2425)) for reproducibility, we obtain: $\boxed{\text{Accuracy} = 0.8355263}$.

Therefore we can conclude that the best criterion is the second one ($\alpha = 0.65$). Based on this value of α, the corresponding confusion matrix for the logistic regression model is given by Table 4.2.

Table 4.2: Confusion Matrix for Logistic Regression

		True	
		Malignant	Benign
Predicted	Malignant	105	6
	Benign	12	29

Once the confusion matrix is computed, we can calculate the sensitivity and the specificity as given by (3.27) and (3.28), respectively. The values that we obtain for this criterion are:

$$\boxed{\text{Sensitivity} = 0.8974359}$$

$$\boxed{\text{Specificity} = 0.8285714}$$

The sensitivity is higher than the specificity for the logistic regression model, which is desirable.

Next, the area under the ROC curve (AUC-ROC curve) can be computed. Taking a sequence of values for α in the interval $[0, 1]$, we can calculate the different values taken by the sensitivity and the specificity. Plotting these values we obtain the ROC curve of the logistic regression method as given by Figure 4.5. The area under the curve of Figure 4.5 is:

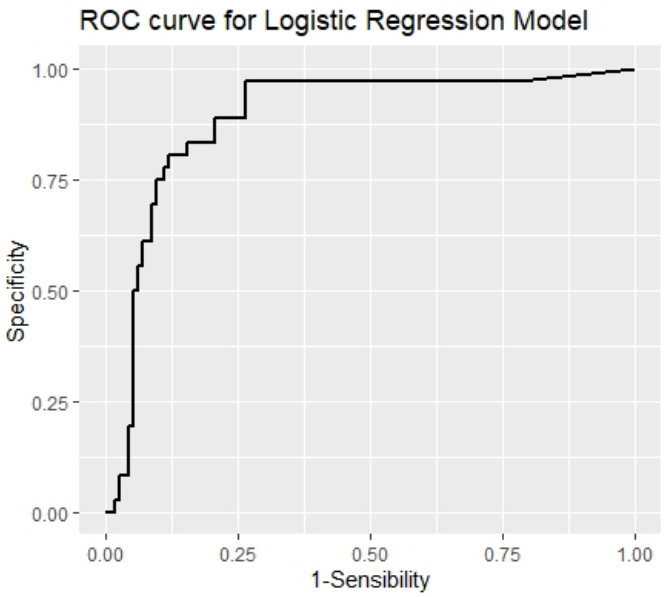


Figure 4.5: ROC curve for the Logistic regression model.

$$\boxed{\text{AUC logistic regression} = 0.9124542}$$

Finally we will calculate the values of the statistics described in (3.10) (3.11), to obtain goodness of fit test. To calculate the likelihoods involved we use the formula (3.5) evaluated in the train set that we used to develop the model. The results that we obtain are:

$$\boxed{D_{null} - D_{fitted} = 3084.584}$$

$$\boxed{R^2_{McF} = 0.931683}$$

We can observe that the deviance value is very large and the pseudo R squared statistic is close to 1. Therefore we conclude that the model developed throughout this section is a good model for fitting our data, and for making predictions.

4.3 Support vector machine

This section contains the performance evaluation of the different SVM models discussed in Section 3.3. First, we will look at the variable importance in Subsection 4.3.1. We can only do this for the linear SVM, and it is carried out only for this model for this reason. Next, we look at the actual performance analysis of the three different SVM methods (linear (Subsection 4.3.2), polynomial kernel (Subsection 4.3.3), and Gaussian kernel (Subsection 4.3.4)). Finally, we will compare these three methods and decide which method works best for our data.

4.3.1 Variable importance SVM

Firstly, we look at the variable importance as given by Figure 4.6

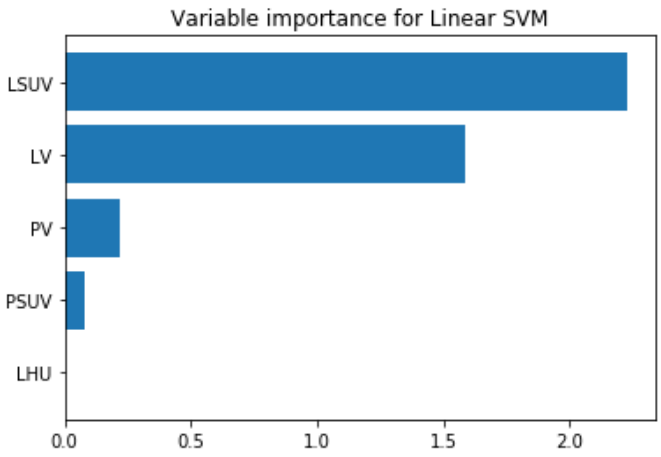


Figure 4.6: Variable importance for Linear SVM

We can easily conclude that LSUV (the mean SUV value of the lymph node) has the most importance for linear SVM. A close second is LV (the volume of the lymph node). The rest of the variables are way less important.

4.3.2 Linear SVM

We investigate for which value of C linear SVM gives the best result. We will compare the accuracy (the fraction of correctly classified data points) to do this, see Table 4.3.

Table 4.3: Accuracy for different values of C

C	0.1	1	10	100	1000
Accuracy	0.822	0.875	0.875	0.875	0.875

We get the best accuracy if $C = 1$ or higher. Applying this model on the test data gave the following results:

$$\boxed{\text{Accuracy} = 0.875}$$

$$\boxed{\text{AUC score} = 0.904}$$

$$\boxed{\text{Sensitivity} = 0.906}$$

$$\boxed{\text{Specificity} = 0.771}$$

And the confusion matrix given in Table 4.4.

Table 4.4: Confusion matrix for linear SVM

		True	
		Malignant	Benign
Predicted	Malignant	106	8
	Benign	11	27

4.3.3 Polynomial kernel SVM

We want to find the optimal parameters for SVM using a polynomial kernel. After looking at some results we noticed that changing the coefficient (r) and γ was less important for this model. That is the reason why we looked at different accuracy for different degrees and values of C (with $\gamma = r = 1$), see Table 4.5.

Table 4.5: Accuracy of SVM with a polynomial kernel for different degrees and values C.

degree\C	0.1	1	10	100	1000
1	0.822	0.875	0.875	0.875	0.875
2	0.842	0.822	0.836	0.829	0.829
3	0.829	0.862	0.875	0.868	0.849
4	0.862	0.862	0.842	0.855	0.842
5	0.829	0.836	0.809	0.829	0.829
6	0.816	0.783	0.803	0.803	0.803

We get the best accuracy when using $C = 10$ and a degree of 1 (which is actually a linear SVM) or a degree of 3. We decided to look into the $C = 10$ and degree 3, because we already looked at the linear case. We checked if we were able to improve the accuracy by changing the coefficient and γ.

Table 4.6: Accuracy of SVM with a polynomial kernel for different degrees and values C.

$r\backslash\gamma$	0.1	1	10
0.1	0.829	0.849	0.862
1	0.855	0.875	0.842
10	0.855	0.888	0.822

From Table 4.6 we get that it is indeed best to have γ set to one, but we also get that we have an better accuracy (0.888) if we let the coefficient be 10.

Using $C = 10$, degree $= 3$, $\gamma = 1$ and $r = 10$ we get the following results after applying the model on the test data:

$$\boxed{\text{Accuracy} = 0.888}$$

$$\boxed{\text{AUC score} = 0.967}$$

$$\boxed{\text{Sensitivity} = 0.915}$$

$$\boxed{\text{Specificity} = 0.800}$$

And the confusion matrix given in Table 4.7.

Table 4.7: Confusion matrix of SVM with a polynomial kernel.

		True	
		Malignant	Benign
Predicted	Malignant	107	7
	Benign	10	28

The accuracy of SVM with a polynomial kernel is higher than the accuracy of linear SVM, just as the AUC score, sensitivity and specificity. From the confusion matrices we get that for both the benign and malignant lymph nodes the number of good classifications increased by one. Combining this, we can conclude that we prefer SVM with a polynomial kernel over linear SVM.

4.3.4 Gaussian kernel SVM

We want to find the optimal parameters for SVM using a Gaussian kernel. After a general search, see Table 4.8, we see that we get the best accuracy if both γ and C are between 1 and 10.

Table 4.8: Accuracy of SVM with a Gaussian kernal for different values of γ and C.

γ\C	0.1	1	10	100	1000
0.1	0.770	0.822	0.862	0.895	0.908
1	0.770	0.908	0.882	0.868	0.882
10	0.770	0.829	0.822	0.822	0.822
100	0.770	0.770	0.783	0.783	0.783

We get the best accuracy (0.921) for $C = 5$ and $\gamma = 1$, see Table 4.9.

Table 4.9: Accuracy of SVM with a Gaussian kernel for different values of γ and C.

γ\C	1	2	3	4	5	6	7	8	9	10
1	0.908	0.908	0.908	0.914	0.921	0.914	0.901	0.895	0.888	0.882
2	0.888	0.914	0.895	0.888	0.888	0.888	0.868	0.855	0.849	0.849
3	0.888	0.895	0.888	0.882	0.875	0.862	0.862	0.862	0.855	0.862
4	0.862	0.862	0.862	0.862	0.849	0.849	0.862	0.855	0.862	0.862
5	0.849	0.868	0.862	0.855	0.855	0.862	0.862	0.855	0.855	0.855
6	0.849	0.842	0.836	0.829	0.829	0.829	0.829	0.829	0.829	0.829
7	0.842	0.836	0.836	0.836	0.829	0.829	0.829	0.829	0.829	0.829
8	0.842	0.829	0.829	0.829	0.829	0.822	0.822	0.822	0.822	0.822
9	0.836	0.829	0.829	0.829	0.829	0.829	0.829	0.829	0.829	0.829
10	0.829	0.816	0.822	0.822	0.822	0.822	0.822	0.822	0.822	0.822

We get the following results for these parameters after applying the derived model on the test data:

$$\boxed{\text{Accuracy} = 0.921}$$

$$\boxed{\text{AUC score} = 0.966}$$

$$\boxed{\text{Sensitivity} = 0.957}$$

$$\boxed{\text{Specificity} = 0.800}$$

And the confusion matrix given in Table 4.10.

Table 4.10: Confusion matrix for SVM with a Gaussian kernel

		True	
		Malignant	Benign
Predicted	Malignant	112	7
	Benign	5	28

The accuracy of SVM with a Gaussian kernel is higher than the accuracy of SVM with a polynomial kernel, just as the AUC score and sensitivity. From the confusion matrices we get that the number of good classifications of benign lymph nodes does not change. However, more malignant lymph nodes get a good classification.

The specificity does not increase. This is less important to us, because it is more important to classify all malignant lymph nodes correct than classifying all benign lymph nodes correct.

From this we conclude that for this model it is best to use a Gaussian kernel when we use SVM, because it has an accuracy of 0.921 and an higher sensitivity than the linear and polynomial SVM's.

4.4 Decision trees

In this section, the performance of the decision tree methods is assessed. First, the variable analysis is carried out in Subsection 4.4.1. Next, the actual performance of the decision tree, bagging and random forest is done in Subsection 4.4.2.

4.4.1 Variable analysis

We start with comparing the variable importance for the three different methods. The variable importance of the decision tree is visualised in Figure 4.7.

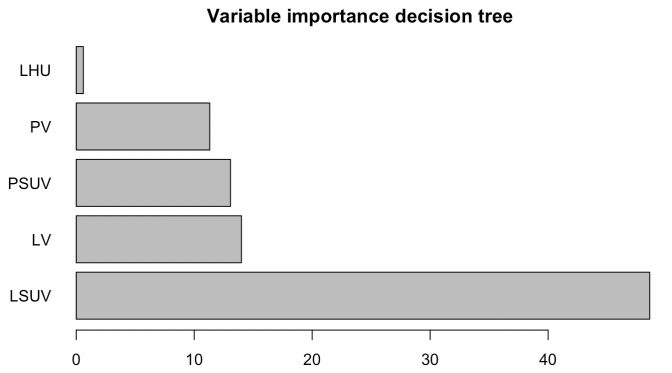


Figure 4.7: Variable importance plot for decision trees

From figure Figure 4.7, it can be seen that the LSUV value has the most importance for this model, it has the most influence on the model.

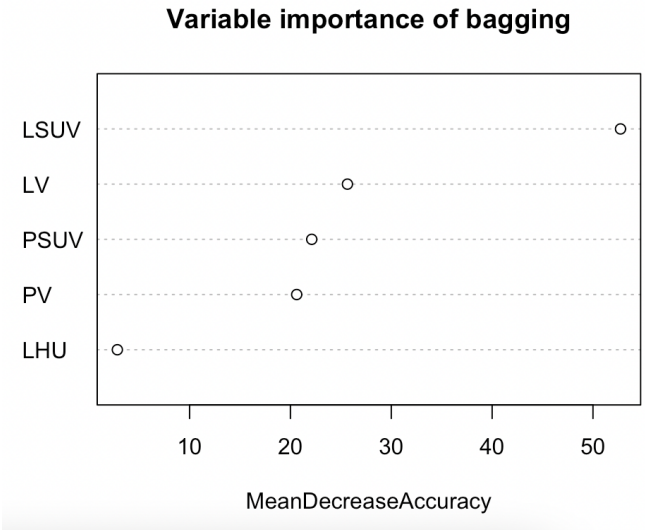


Figure 4.8: Variable importance plot for bagging

Also in the variable importance plot for bagging, Figure 4.8 we see that the LSUV variable has the most importance to the model. All the other variables almost have the same ranking compared to the variable importance of decision trees, except LHU, which is the least important.

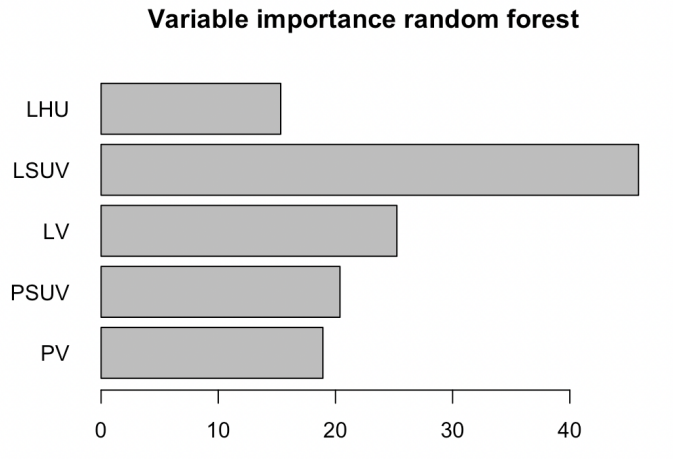


Figure 4.9: Variable importance for random forest

In the variable importance plot of random forest, as given by Figure 4.9, we also have the result that LSUV has the most impact on the model. Also here is the LHU variable the least important. However, the LHU variable has an higher variable importance compared to the variable importance of the pruned decision tree and the bagging method.

4.4.2 Performance analysis

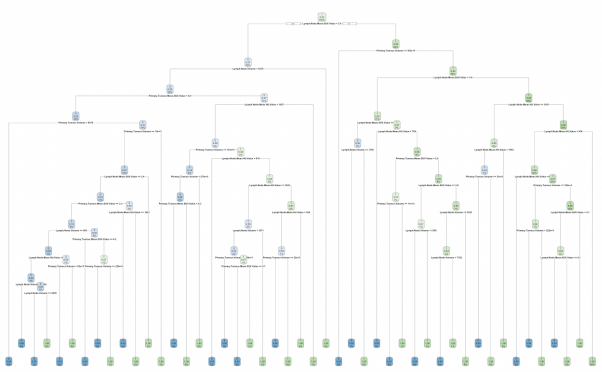
In this subsection, the performance analysis is carried out for the three decision tree methods. First, the pruned decision tree is evaluated, next the bagging method is evaluated, and finally the random forest method is evaluated.

Decision trees

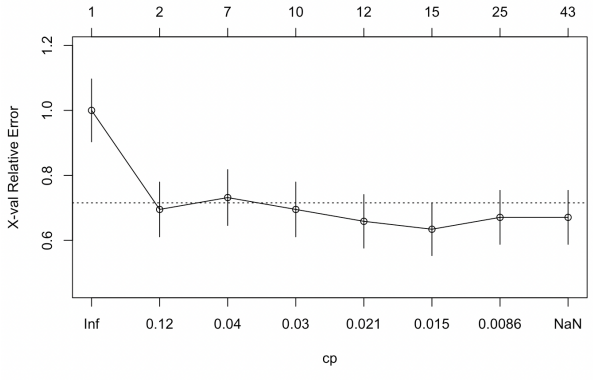
We start with applying the decision tree on the whole dataset. For this, we also plot the complexity parameter plot which will help us determine the best size for our tree. Growing the full tree gives us Figure 4.10a.

As we can see, this tree becomes very complex and tends to overfit the data, which gives poor generalisation results. By using the 1-standard-error [2] rule, it can be seen from Figure 4.10b that from size 7-10 onwards, the line flattens out. This means that a tree with 7-10 splittings gives the best result. When using the default `rpart` function in R, it chooses automatically the best size for our tree. When doing this, we get the result from figure Figure 4.11a.

[2] https://bradleyboehmke.github.io/HOML/DT.html [Last Accessed 26-05-2021]

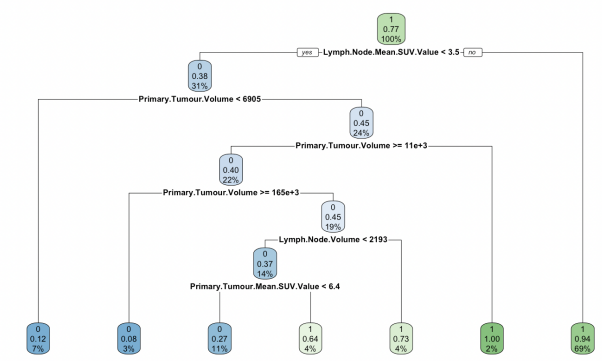


(a) Fully grown classification tree of the data

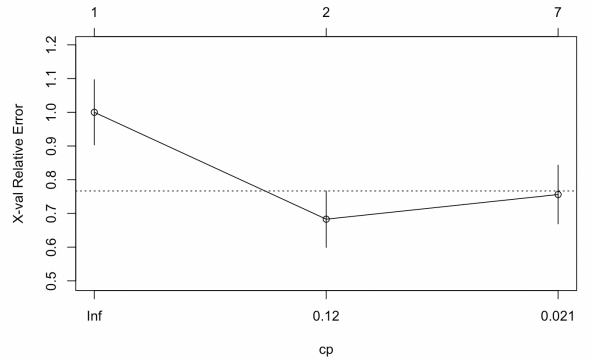


(b) Pruning cp plot for the fully grown classification tree

Figure 4.10: Results of the fully grown classification tree



(a) Automated classification tree of the data



(b) Pruning cp plot for the classification tree

Figure 4.11: Results of the automated classification tree

We see that this tree has 6 splits and 7 terminal nodes. This shows that the default setting almost has the best performance.

In Table 4.11 the confusion matrix is shown for the pruned decision tree.

Table 4.11: Confusion matrix for decision trees

		True	
		Malignant	Benign
Predicted	Malignant	109	11
	Benign	8	24

We see that there are 24 true positives and 109 true negative. Also, there are 11 false positives and 8 false negatives. From this we can calculate our accuracy, sensitivity and specificity. We also give the AUC.

$$\boxed{\text{AUC pruned decision tree} = 0.809}$$

$$\boxed{\text{Accuracy pruned decision tree} = 0.875}$$

$$\boxed{\text{Sensitivity pruned decision tree} = 0.932}$$

$$\boxed{\text{Specificity pruned decision tree} = 0.686}$$

Bagging

First, we will compare different tree sizes for the bagging method. We plot for 0-500 trees the AUC, we want our AUC to be as high as possible. This can be seen in Figure 4.12.

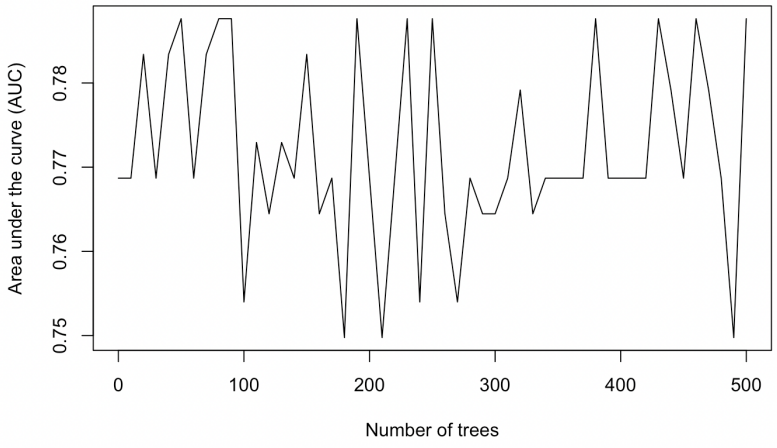


Figure 4.12: Bagging different amount of trees vs the AUC

We see that the AUC differs per number of trees and there is no clear stationary result. Therefore, for the further results we will use 200 as the number of trees.

When using bagging, it gives the confusion matrix given by Table 4.12.

Table 4.12: Confusion matrix for bagging

		True	
		Malignant	Benign
Predicted	Malignant	108	9
	Benign	9	26

We also calculated here the measures:

$$\boxed{\text{AUC bagging} = 0.833}$$

$$\boxed{\text{Accuracy bagging} = 0.882}$$

$$\boxed{\text{Sensitivity bagging} = 0.923}$$

$$\boxed{\text{Specificity bagging} = 0.743}$$

Random forest

For our random forest, we get the following results. The confusion matrix is given by Table 4.13.

Table 4.13: Confusion matrix for random forest

		True	
		Malignant	Benign
Predicted	Malignant	109	10
	Benign	8	25

The rest of the measures are:

$$\boxed{\text{AUC random forest} = 0.823}$$

$$\boxed{\text{Accuracy random forest} = 0.882}$$

$$\boxed{\text{Sensitivity random forest} = 0.932}$$

$$\boxed{\text{Specificity random forest} = 0.714}$$

Summary of peformance evaluation

Table 4.14: Results for the 3 different methods

Method	**AUC**	**Accuracy**	**Sensitivity**	**Specificity**
Pruned decision tree	0.809	0.875	0.932	0.686
Bagging	0.833	0.882	0.923	0.743
Random forest	0.823	0.882	0.932	0.714

In Table 4.14 are all the results for the 3 methods we used in this section. If we look at the AUC, the bagging model performed the best. The accuracy is the highest for the random forest method and the bagging model. The sensitivity is the highest for the pruned decision tree and the random forest method. The specificity is not that good for all methods, the bagging model preformed the best. In general, one would expect that random forest gives better results. This also followed from our robustness analysis. A reason for the fact that the bagging model in our case has better performance, is probably just luck with the specific data set. Increasing the number of trees does not necessarily influence the performance of the model. This can be seen in Figure 4.12. This means that as the random forest model is more robust, it is recommended to use that model for this type of problem.

In conclusion, the decision tree model has good performance and especially bagging and random forest show good improvements compared to the thresholding model.

4.5 Comparing the hospitals

Because Gaussian kernel SVM performed the best out all models, we used this method to perform an additional assessment on whether there are major differences between the data sets from the different hospitals (Albert Schweitzer Hospital (ASZ) and the Diakonessen Hospital (DIAK)). We used the data from the ASZ to train our model and we used the data from the DIAK to test this model. In case the performance of the model does not differ greatly with respect to the previously presented results, this would mean the data from the different hospitals is similar. Tuned hyperparameters where found in the same way as in Subsection 4.3.4, which resulted in $\gamma = 0.06$ and $C = 100$. We get the following results.

$$\boxed{\text{Accuracy} = 0.888}$$

$$\boxed{\text{AUC score} = 0.953}$$

$$\boxed{\text{Sensitivity} = 0.920}$$

$$\boxed{\text{Specificity} = 0.783}$$

And the confusion matrix given in Table 4.15.

Table 4.15: Confusion matrix of SVM with a polynomial kernel.

		True	
		Malignant	Benign
Predicted	Malignant	69	5
	Benign	6	18

Though slightly worse, these results are almost similar to the results we get using a mixed training and test set. The boxplots of Figure 2.8 already foreshadowed this result by showing that the data did not differ greatly. Combined with the analysis performed in Subsection 2.4.5, it can therefore be concluded that there does not exist a significant difference between the data sets from the different hospitals.

4.6 Robustness analysis

The models presented in the previous subsection are all trained and evaluated on the same train and test data set. This was done in order to achieve an equal performance analysis for all models. However, during model performance evaluation it was observed that a change in train and test data may influence the performance of the models, where it may have a strong influence on one model and a weak influence on another. From this observation, the need of a robustness analysis arises. In order to assess the change in accuracy with a change in train and test data, a robustness analysis is carried out in this section. The data set is randomly split into train and test data according to the procedure mentioned in Subsection 2.4.6 for 100 runs, meaning each run contains a different train and test data set. The accuracy of the logistic regression model (Section 4.2), all SVM models (Section 4.3), and all decision tree models (Section 4.4) is then computed for each run from, resulting in an accuracy distribution which forms the basis of the robustness analysis.

The accuracy distribution of the thresholding model, the logistic regression model, all SVM models, and all decision tree models are visualised in the form of kernel density estimations (KDE) in Figure 4.13.

Additional to the aforementioned visualisations, the mean and variance of the distributions visualised in the previous figures are tabulated for all models in Table 4.16.

Table 4.16: Accuracy mean and variance for all models according to the robustness analysis.

(a) Thresholding.

	Thresholding
Accuracy mean	0.8369
Accuracy variance	6.384×10^{-4}

(b) Logistic regression.

	Logistic Regression
Accuracy mean	0.8556
Accuracy variance	5.333×10^{-4}

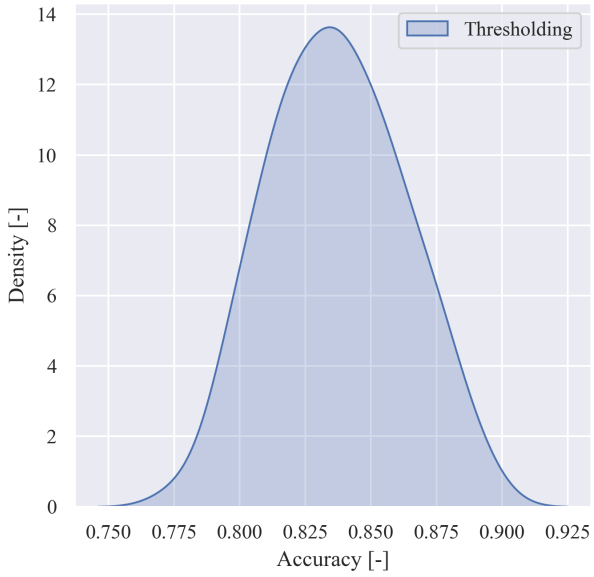
(c) SVM.

	Linear	**Polynomial Kernel**	**Gaussian Kernel**
Accuracy mean	0.8632	0.8768	0.8866
Accuracy variance	4.960×10^{-4}	4.329×10^{-4}	3.910×10^{-4}

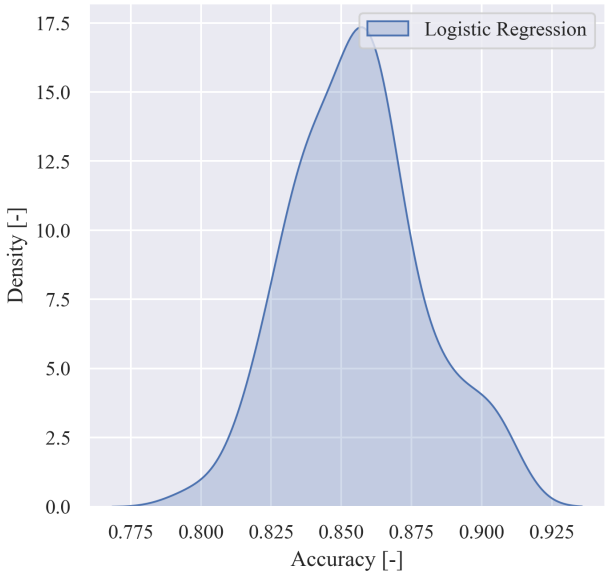
(d) Decision trees.

	Decision tree	**Bagging**	**Random forest**
Accuracy mean	0.8439	0.8655	0.8722
Accuracy variance	4.722×10^{-4}	5.992×10^{-4}	4.399×10^{-4}

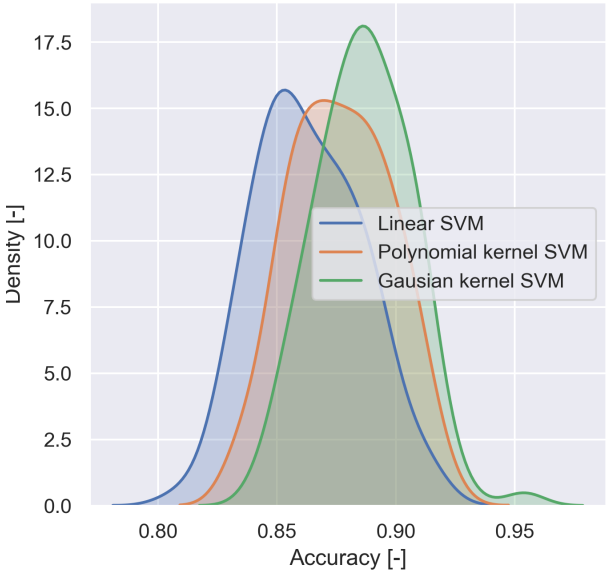
From Figure 4.13 and Table 4.16, it can be seen that the mean and the variance of the accuracy scores lie relatively close to each other for most models. Considering its simplicity, the thresholding model performs surprisingly well, with a mean that competes with the other models but is slightly lower, and an only slightly higher variance.



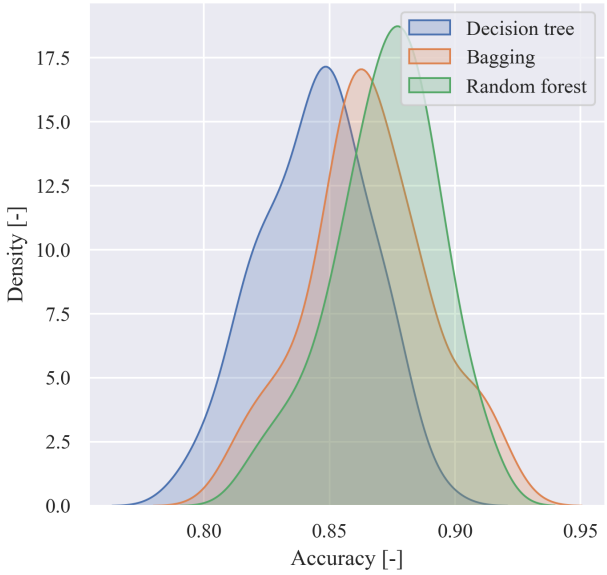
(a) Thresholding.



(b) Logistic regression.



(c) SVM.



(d) Decision trees.

Figure 4.13: KDE of model accuracy score distributions.

The mean accuracies for all the methods do not deviate by much from the results presented based on the single train and test set. Together with relatively low variance, it can therefore be stated that these models are robust enough to apply on a data set of similar sizes compared to the data set used in this analysis (see Chapter 2).

5 Conclusion

This report detailed the construction and evaluation of several different models that classify lymph nodes to be either benign or malignant based on the image data features listed in Section 2.3. A conclusion is now drawn based on these findings. The original research question was stated as follows:

- "Is it possible to predict if a lymph-node is benign or malignant from the properties of the 3D segmented lymph-node and the primary tumour of the patient?"

With the additional sub-research questions to be answered:

1. "How much can the predictive performance be increased with respect to the simple thresholding model proposed by the Albert Schweitzer Hospital?"
2. "Does the information from the primary tumours help to improve lymph node classification predictions?"
3. "Are there differences between the datasets of the two different hospitals?"
4. "Does the CT data correlate with the classification of the lymph nodes and if so in what respect?"

Table 5.1 provides an overview of the results from the different assessments of Chapter 4. The methods are ordered based on their accuracy. If two methods had the same accuracy we looked at the sensitivity for the reasons mentioned in Section 3.5.

Table 5.1: Summary of the results for the different methods

Method	**Accuracy**	**AUC**	**Sensitivity**	**Specificity**
Gaussian kernel SVM	0.921	0.966	0.957	0.800
Polynomial kernel SVM	0.888	0.967	0.915	0.800
Random forest	0.882	0.823	0.932	0.714
Bagging	0.882	0.833	0.923	0.743
Logistic Regression	0.881	0.912	0.897	0.829
Pruned decision tree	0.875	0.809	0.932	0.686
Linear SVM	0.875	0.904	0.906	0.771
Thresholding	0.849	0.896	0.855	0.829

The conclusion is structured as follows. First, some general remarks on the results of Table 5.1 are made in Section 5.1. Next, some more specific conclusions are drawn on each of the models, starting with the thresholding model in Section 5.2, followed by the logistic regression model in Section 5.3. Continuing, the conclusions on the support vector machines models are presented in Section 5.4 and finally for decision trees in Section 5.5.

5.1 General remarks

In general, we notice that all models perform well, and better than the 70% to 80% accuracy of nuclear medicine physicians. This directly answers the core research question of this report. Based on the obtained accuracies we conclude that it is in fact possible to predict if a lymph-node is benign or malignant from the properties of the 3D segmented lymph-node and the primary tumour of the patient. Enough information is contained in the image data to obtain a maximum accuracy of 92%. This means that models such as the ones constructed in this report could potentially serve complementary to the judgement of a nuclear medicine physician. Furthermore, all methods are at better than the simple thresholding model, which also already performed well with an accuracy of 0.849 and a sensitivity of 0.855. This directly answers sub-research question 1. The more complex models can in fact perform better than the simple thresholding model provided by the Albert Schweitzer Hospital.

Secondly, for all methods the specificity is (significantly) lower than the sensitivity. This is not an extreme issue though, because in practice this means that more malignant lymph nodes are correctly classified than benign lymph nodes, as explained in Section 3.5.

We can conclude that Gaussian kernel SVM performs the best. This method has both the best accuracy and sensitivity. Furthermore, from Section 4.6 we get that this method has the highest accuracy mean and the lowest accuracy variance, meaning this model is also the most robust. The specificity of this model also ranks second highest.

Concerning the variable analysis; for all the models, the lymph node mean SUV is the most important variable. All the other variables have significantly less influence, but still some influence on the predictions with the different models. This means the primary tumour data does help to classify a lymph node, but the data about the lymph node is the most important. This answers sub-research question 2, and is further elaborated upon in the upcoming subsections. The CT data (the LHU variable) was found not to be significantly important to get a good classification, which poses an initial answer to sub-research question 4.

Concerning the data provided by the different hospitals, the boxplot analysis and the Gaussian SVM analysis showed that there is not a significant difference between the data provided by the Albert Schweitzer Hospital and the Diakonessen Hospital. This directly answers sub-research question 3. Small differences in the data set could be explained from the fact that the data set provided by the Diakonessen Hospital is extremely small.

5.2 Thresholding

With regards to the thresholding model proposed by the Albert Schweitzer Hospital, it can be observed that it performs reasonably well considering its (lack of) complexity. Together with the logistic regression model, it even scores highest in terms of specificity. However, as explained earlier, sensitivity and accuracy are more important measures than specificity, and especially in terms of sensitivity

the thresholding model under-performs with respect to the more complex models. However, the fact that the thresholding model performs so well leads to the conclusion that the CT scan data and primary tumour data are not as important to the classification as initially expected. Chapter 1 mentioned that nuclear medicine physicians can not determine from just looking at the lymph node only whether it is benign or malignant, but that the shape and intensity of the primary tumour influences their decision. The lowest accuracy of the thresholding model confirms this expectation, though it proves that the improvements of including the primary tumour data are only marginal. Based on these findings, sub-research question 2 and 4 can partially be answered. The information from the primary tumour may help to improve the lymph node classification predictions, though these improvements are only marginal and most of the information does in fact seem to be contained in the lymph node mean SUV value. Furthermore, the CT data may correlate with the classification of the lymph nodes. However, once again this correlation is small and may only lead to marginal improvements. From the thresholding model only, it can not be seen whether the increase in performance of the other models is due to the inclusion of primary tumour data or the CT data, or both.

5.3 Logistic regression

The logistic regression model performs reasonably well in terms of accuracy, ranking 5^{th} of all models, though the scores of the random forest and bagging are extremely close. A downside is its low sensitivity. Not considering the thresholding model, the logistic regression model ranks lowest in terms of sensitivity out of all models. The specificity is high however, though it has already been discussed that this measure is not as important as sensitivity. As such, it could be concluded based on the sensitivity specificity ratio that the logistic regression model is not perfect for the application of medical classifications.

It was mentioned in Subsection 3.2.5 that the model assumes independent predictors. Since Figure 2.7 showed that there exists some correlation between certain predictors, this assumption may be invalid. However, judging by the reasonable accuracy, it could be concluded that the correlation between predictors is small enough to validate the assumption of independent predictors. Being aware of this limitation, the model performance could possibly be increased by the use of generalised additive models to allow non-linear dependence of the predictors Hastie et al. (2001). This does not guarantee improvements, since this can lead to overfitting the training data set, but it is a possible recommendation to improve model performance.

The variable importance analysis for logistic regression showed the variables related to the primary tumour and the CT data to be rather statistically insignificant. This once again answers the sub-research questions related to the inclusion of this data in the predictions; sub-research question 2 and 4. Namely, they do not contribute greatly to the predictive accuracy. This means that another possible way to improve the logistic regression model is by p-norm regularisation with $p < 2$, since this will eliminate insignificant predictor variables completely and will introduce sparsity. However, since there is already a small number of predictor variables, eliminating some of them by p-norm regularisation could lead to instabilities in our model, so it would have to be carefully assessed.

In line with the above, we could think about introducing many more predictors to our model, or introduce only the 2 most important predictors. However, the ratio between the number of predictors selected and the magnitude of our data set was considered to be adequate [1]. Increasing the number too much or decreasing it could affect the stability of our model, worsening the obtained results.

5.4 Support vector machines

In general we can assume that an SVM classifier performs well enough to predict if lymph-nodes are benign or malignant. From the three SVM classifiers we looked into, the Gaussian kernel SVM performed the best. This SVM gave the highest accuracy, AUC score and sensitivity compared to the other models and the specificity was relatively high as well. The linear SVM performed the least good, compared to the other Gaussian and polynomial kernel SVM. This is as we expected, because using a polynomial kernel of degree one is the same as linear SVM and also the Gaussian kernel can be tuned in such a way that it performs similarly as the linear SVM. Hence, polynomial and Gaussian kernel SVM will always perform the same or better as linear SVM, dependent on the chosen hyper parameters. Since we chose hyper parameters that creates polynomial and Gaussian SVM's different than the linear SVM, we may conclude that the given dataset is not linearly separable. This can be taken into account when further investigating the models.

5.5 Decision trees

For decision trees, we see that the ensemble methods give the best results. This includes both bagging and random forest. For the general training set used in this report, bagging gave better results. Especially this model provided high accuracy and sensitivity. However, the random forest model is more robust. With an accuracy of 87%, this model is one of the better models for this kind of research. In general, you would expect that random forest gives better results. This also followed from our robustness analysis. A reason for the fact that the bagging model in our case has better performance, is probably just luck. Increasing the number of trees does not necessarily influence the performance of the model.

Fully grown decision trees and pruned trees gave a lower accuracy which made those models less attractive. In conclusion, for this kind of classification trees, it is recommended to use bagging or random forest.

[1] Dr. N. Parolya - Personal communication [Last Accessed 14-05-2021]

6 Discussion

This chapter lists some points of discussion that have originated from the analysis of this report. Some of these discussion points lead to recommendation for future research.

The number of lymph nodes per primary tumour per patient in the data set was too small in order to classify the lymph nodes based on there locations in the mediastinum. For further research it would be ideal expand the data set such that this can be included as a feature, because the location of the lymph node with respect to the primary could affect the probability of it being malignant, since cancer spreads from the primary tumour towards the lymph nodes, lymph nodes that are located closer to the primary tumour may have a higher chance of being malignant compared to those located further away. As such, it would be interesting to look at the location of the tumour in relation with the location of the lymph node.

The application of the CT scan data to the models may not perform as well as expected. This due to the limitation of the CT scan not being perfectly aligned with the PET scan due to the presence of air in the mediastinum. This means that the masks which are initially created based on segmentation of the PET-CT scan images, may be offset when applied to the CT scan images. As a result, the inclusion of CT data may compromise the performance of the classification models since the masks are not properly aligned. Different regions of the mediastinum may be investigated for the PET-CT and CT scan data by the models due to this. In case the masks do not align properly, this may be a possible reason for the LHU variable being the least important variable in almost all models.

As mentioned in Section 4.6, the performance of the models depends on how the random seed was set when splitting the data into a train and test set. Though the models were shown to be robust, one way to account for this effect is to train and test the models using K-fold cross-validation. Furthermore, the change in accuracy with random seed when splitting the data set is expected to decrease and the robustness of the models is expected to increase when trained on a larger data set, since the data set provided in this analysis was rather small. Therefore, it would be beneficial to perform a study of the models on a significantly larger data set. This would also allow the use of even more complicated models such as deep learning neural networks (though complications regarding different image sizes due to the applied masks would have to be resolved as well in order for this to be applicable).

With regards to the previous discussion item, the size of the data set could be increased by introducing image augmentation. Specifically, additional data can be created by creating rotations of the existing images and adding them to the data set. Deforming or shape shifting the images is not recommended however, since this may lead to unrealistic lymph node and primary tumour shapes and volumes that do not occur in practice.

In the robustness analysis, the hyperparameters of the SVM models were not tuned to the extent they were tuned in Section 4.3. As a result, most likely, the mean accuracy of the SVM models will be higher and the variance will be lower. This should be taken into account when judging the results of the robustness analysis. In order to perform a complete robustness analysis, the hyperparameters need to be tuned to the extent of Section 4.3. Though extremely computationally and time intensive, this is recommended if the true performance of the SVM models needs to be assessed.

References

Domencich, T. A. & McFadden, D. L. (1975), *Urban Travel Demand: A Behavioral Analysis*, North-Holland, Amsterdam, The Netherlands.

Geron, A. (2017), *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*, O'Reilly Media, Sebastopol, CA.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The Elements of Statistical Learning*, Springer Series in Statistics, Springer New York Inc., New York, NY, USA.

Hosmer, D. W. & Lemeshow, S. (2000), *Applied logistic regression*, John Wiley and Sons.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013), *An Introduction to Statistical Learning: with Applications in R*, Springer.
URL: *https://faculty.marshall.usc.edu/gareth-james/ISL/*

Kalender, W. A. (2011), *Computed Tomography: Fundamentals, System Technology, Image Quality, Applications*, Wiley.
URL: *https://books.google.nl/books?id=gfLWmRjoyPMC*

Longford, N. T. (1987), 'A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects', *Biometrika* **74**(4), 817–827.
URL: *http://www.jstor.org/stable/2336476*

Rawat, Y. (2020), 'Non-linear svm and kernel function'.
URL: *https://medium.datadriveninvestor.com/non-linear-svm-and-kernel-function-7174bbecc2d3*

Stehman, S. V. (1997), 'Selecting and interpreting measures of thematic classification accuracy', *Remote Sensing of Environment* **62**(1), 77–89.
URL: *https://www.sciencedirect.com/science/article/pii/S0034425797000837*