École Polytechnique Fédérale de Lausanne

Predicting Ebola diagnosis and outcome using machine learning: A retrospective cohort study on the 2014-16 West African epidemic

by Ridha Chahed

Semester Project Report

Approved by the Examining Committee:

Mary-Anne Hartley
Project Supervisor

Prof. Martin Jaggi
Project Advisor

EPFL IC MLO
Lausanne, Switzerland

September 30, 2022

Perhaps we should all stop for a moment and focus not only on making our AI better and more successful but also on the benefit of humanity.
— Stephen Hawking

Dedicated to Selim, Kaouther, Hend and my friends.

Acknowledgments

First and foremost I would like to pay tribute to the medical professionals present on the front to fight diseases and save lives.

I would like to express my sincere gratitude to my advisor Dr.Mary-Anne Hartley for her continuous support in this project. Her energy, motivation and vision have deeply inspired me, I couldn't have imagined a better mentor for this project.

Finally, I would like to thank Professor Martin Jaggi for offering me the opportunity to work with his group.

The data were contributed to the IDDO Ebola Data Repository by the following organisations:

- Alliance for International Medical Action (ALIMA)
- International Medical Corps (IMC)
- Institute of Tropical Medicine Antwerp (ITM)
- Médecins Sans Frontières (MSF)
- Oxford University
- Save the Children (SCI)

Lausanne, September 30, 2022 Ridha Chahed

Abstract

Background
In an unprecedented effort to centralize medical data, the Infectious Disease Data Observatory (IDDO) curated the largest clinical dataset on Ebola Virus Disease (EVD). Until now, this neglected emerging disease has been studied on small, fragmented datasets which dilutes statistical power and results in poor generalizability.

Aim
We join IDDO's effort to standardize and analyse this large data set with the aim of improving probabalistic diagnosis and risk stratification for Ebola using machine learning models.

Methods
From the total curated data set of 13562 patients, this study selects 2/12 subsets representing 3370 patients from 3 Ebola treatment centers erected during the 2014-16 West African EVD epidemic. We develop a proof of concept data cleaning pipeline, provide descriptive statistics and build a series of interpretable machine learning (ML) models for diagnosis and risk stratification of EVD. The main ML models compared are logistic regression, SVM, KNN, decision trees, extra trees and random forest. Labels for diagnosis was PCR-confirmed EVD in a blood sample taken at triage, and risk stratification was based on the probability of death and computed only for the EVD+ subgroup. Features used were clinical signs and symptoms as well as laboratory measures taken on day 0.

Findings
Across all subsets of the data, the best diagnostic model had an AUC of **0.84** for diagnosis and **0.85** for prognosis. These results were obtained using ensemble methods such as Random Forest and Extra trees. Excellent results have also been reached with simpler methods such as logistic regression that reached an AUC of **0.77** for diagnosis and **0.84** for prognosis with the benefit of being easier to train.

Conclusion
This work is a template to build a single, standardized and interoperable aggregated dataset on which we can derive and compare local and global predictive models and explore the possibility decentralized learning.

Contents

Chapter 1

Introduction

Ebola Virus disease
The Zaire Ebolavirus appeared in 1976 with 2 simultaneous outbreaks in Nzara, South Sudan and Yambuku, a village near the Ebola River in the Democratic Republic of Congo (DRC) [**3**]. Four decades after its discovery it has killed more than 15 000 people across Africa with a fatality rate averaging 83% [**4**] but still little is known about this disease. The 2014-2016 outbreak is the largest outbreak recorded with more than 28,000 cases and 11,000 deaths. It started in Guinea and crossed over the borders to Sierra Leone and Liberia. Currently there is an ongoing outbreak in the Democratic Republic of Congo where medical coverage is limited by insecurity, poor education and poverty [**5**]. Even with the advent of ring vaccination [**6**], the current epidemic took over a year to control. Ebola spreads within the population via direct contact with body fluid (blood, vomit, feces) of infected persons or objects contaminated by them (fomites). The incubation period varies from 2 to 21 days, transmission is usually only during symptomatic disease and asmyptomatic transmission seems to be rare [**7**].

A diagnostic and prognostic challenge
One major problem of Ebola virus disease (EVD) is that it's hard to distinguish from other diseases like Malaria, typhoid and meningitis [**8**]. Indeed, the majority of patients admitted to Ebola treatment centers for suspected infection did not have Ebola. The same issue exists for prognostic triage. Clincal outcomes range from asymptomatic to fatal and case fatality rates are dependent on their environment. For instance, mortality rates in high resource settings from just 20% compared to 60-80% in West Africa and previous outbreak [**9**].

Improved probabalistic triage for diagnosis and prognosis, would not only limit the risk of nosocomial infections but also better allocate resources to those who need it most, and possibly reduce mortality in resource limited settings. Predictive models are low cost and easy to implement, however many published were based on single study sites with small localised populations, yielding results that are poorly generalizable [**10**].

There is a need to present a general framework to harmonize the data from different studies

in order to compute models. Based on the harmonized data, these models will predict the likely course of a patient's health.

The need for adaptive predictions

After the epidemic, many scoring systems were derived to predict survival outcomes and diagnosis from the various study sites. Using small data sets and simple models such as logistic regression produced encouraging results able to replicate clinician assessment, with an accuracy between 0.64 and 0.74 [**11**]. However, these static models risk becoming irrelevant in changing environments. Indeed, Ebola disease presentation and outcomes was shown to be not only highly heterogeneous across sites, but also in time: where outcomes evolved with emerging strains, changing health care seeking behaviour and health care capacity [**12**].

As machine learning (ML) models are able to learn from incoming data, they represent a promising approach to this issue, whereby predictions can evolve with their environments.

Chapter 2

Aim and Objectives

In this study we aim to create dynamic probabilistic diagnostic and prognostic triage for Ebola Virus Disease using machine learning on a subset of the largest Ebola data set in the world.

1. Develop a data cleaning pipeline and assess the quality of the curated and raw data .
2. Perform descriptive statistics of the local studies and aggregated data
3. Build a series of local ML models for diagnosis and risk stratification of Ebola using on a representative subset of the data
4. Optimize and compare the above models, interpret the results, and assess their generalizability.
5. Summarize the work and identify the next steps.

Chapter 3

Data

3.1 Study design

This retrospective cohort study derives ML models for EVD diagnosis and prognosis.

- **Outcomes.** Diagnostic outcomes are EVD+ or EVD- and based on a PCR test of blood drawn on the day of triage. All patients arriving at triage are included in this model. Prognosis is survival or death of EVD+ patients during their admission at the treatment center.
- **Features.** Features used for prediction are demographics (age, sex) geographic location, date, clinical signs/symptoms collected on day 0 of triage (presence/absence of fever, gastrointestinal complaints, pain etc), contact tracing information and some basic paraclinical tests (malaria rapid test)

3.2 Data sets

All the studies have been centralized by IDDO each comprises individual level data from **13562** individuals across **13** different studies at **16** cities across **3** countries. (See Figure 3.1). Patients ranged in age from 0 to 100 years and were distributed across 3 countries (4471 patients in Sierra Leone, 3623 patients in Liberia and 5468 patients in Guinea).

The studies are summarized below and those selected for use in this project are highlighted in colour.

Table 3.1 – Summary of the cohort origin

Study Identifier	Contributor	Country	City	Number of patients
EQJJGF	Médecins Sans Frontières (MSF)	Liberia	Monrovia	1907
EGOYQN		Guinea	Guéckédou	2500
EJPDEJ			Donka	2301
ERFCVU	International Medical Corps (IMC)	Liberia	Bong	550
			Margibi	292
		Sierra Leone	Port Loko	549
			Kambia	273
			Makeni	1085
EOPNOJ	Alliance for International Medical Action (ALIMA)	Guinea	Nzérékoré	147
EORKWS	Oxford University	Sierra leone	Port Loko	35
EBPOHA		Liberia	Monrovia	4
ESYADD	Save the Children International (SCI)	Sierra leone	Kerry Town	456
ESBMRS	Institute of Tropical Medicine Antwerp	Guinea	Donka	102
EIXUZQ	Médecins Sans Frontières (MSF)	Liberia	Foya	870
EPGLFV		Sierra leone	Freetown	171
EFFVXT	Institute of Tropical Medicine Antwerp	Guinea	Donka	418
EUZJTB	Médecins Sans Frontières (MSF)	Sierra leone	Bo	524
			Kailahun	1219
			Magburaka	159
12 Studies	6 providers	3 countries	16 cities	13562 patients

In this study, we prepare 2 datasets for analysis, consisting of 3370 patients (25 % of the entire dataset). These datasets were selected due to their large size, accessible data dictionary and larger proportion of EVD+ patients (more balanced diagnostic label distribution).

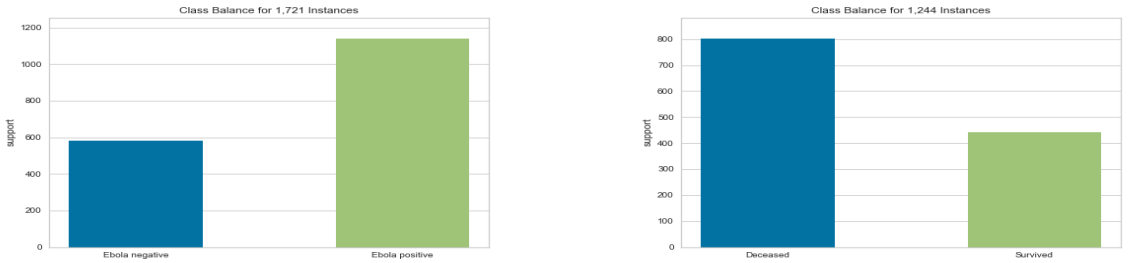
To be concise, **only the models developed for the largest dataset will be presented in this report (EGOYQN)**. However, in a final section we will present some key results obtained with the other study for comparative analysis.

3.3 The EGOYQN study

This study was undertaken by Médecins Sans Frontières (MSF) in Guinea from January 2014 to November 2015 with 2500 patients among which 1372 patients are EVD positive. In this dataset we have at our disposal demographic, temporal and geographic data, clinical signs, symptoms and outcomes (See [**13**] for descriptive analysis).

3.3.1 Outcomes

- Diagnostic outcomes: By only keeping the patients with a confirmed final diagnosis, 1721 patients with 1141 patients EVD positive (1141 / 1721 = 0.66)
- Prognosis labels : By only keeping the Ebola positive patients with a confirmed final vital status we get 1244 patients with 803 patients deceased (803 / 1244 = 0.65)



(a) Diagnosis data

(b) Prognosis data

Figure 3.1 – Outcome distribution

3.3.2 Geographic distribution

The patients are admitted at two centers in Macenta (379 /2500 = 0.15) and Guéckédou (n=1244/2500 = 0.49), additionally, 387 patients were first admitted in Macenta's center and then transferred to Guéckédou. The admission center for 490 patients is unknown. To understand the evolution of the admission see Figure 3.2.

A majority of patients reside in Macenta (969 / 2500 = 0.39) and Guéckédou (656 / 2500 = 0.26) followed by Nzérékoré (271 / 2500 = 0.11), Kérouane (194 / 2500 = 0.08), Kissidougou (107 / 2500 = 0.04) and other prefectures.

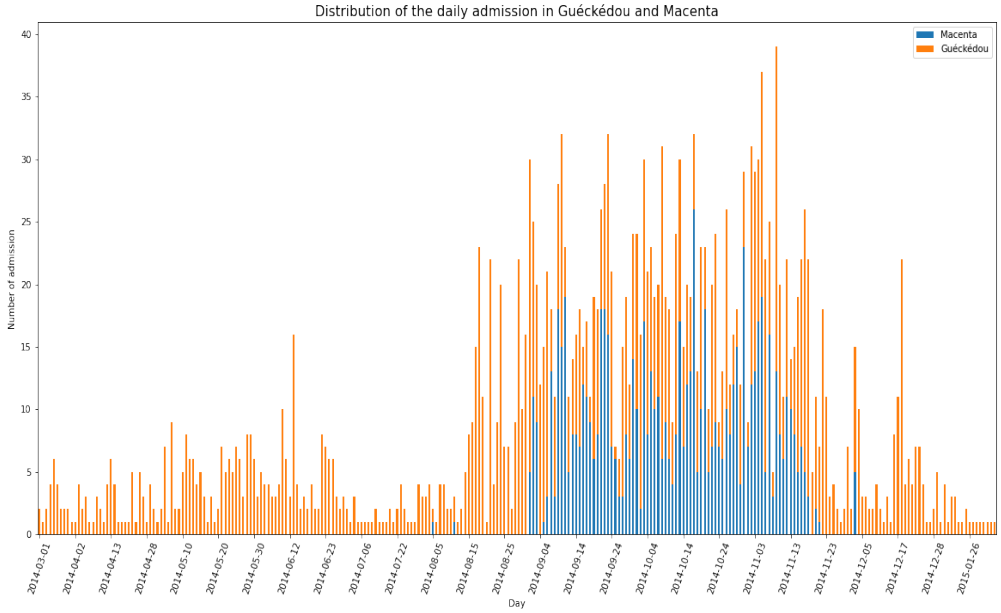


Figure 3.2 – Distribution of the admission day

3.3.3 Demographic composition

Sex. A small majority (1303 / 2500 = 0.52) are females, (1125 / 2500 = 0.45) are males and for 72 patients the sex is unknown.

Occupation. Although 3/5 of the job titles are unknown, house workers (428 / 2500 = 0.17) and farmers (189 / 2500 = 0.08) are particularly frequent.

Age. The age is unknown for 70 patients and the median age is 30 years.

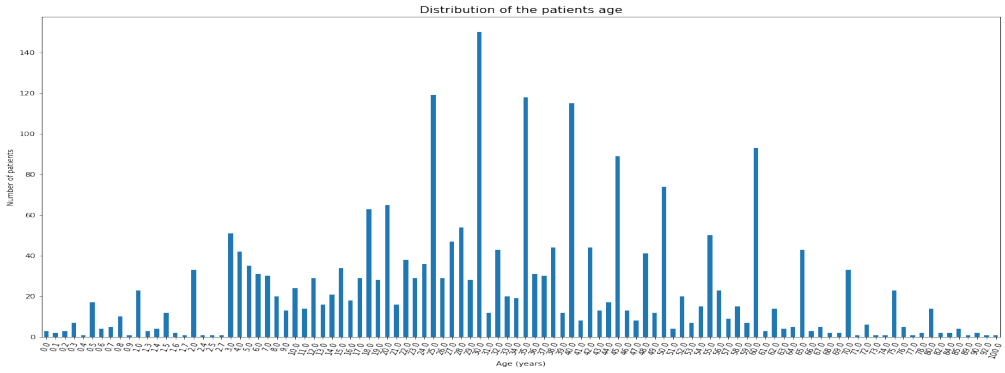


Figure 3.3 – Distribution of age

3.3.4 Dealing with missing data

Removing data points

Due to missing features: We exclude 612/2500 (25%) patients from the cohort due to absent triage symptomatology. Symptom distribution for the remaining 1888 patients is tabulated below according to missingness (see Table 3.2). The group 1 symptoms have almost no missing values. Excluding 69 (3%) patients results in features with no missing values in this group. Thus 1819/2500 (73%) patients remain.

Due to missing outcomes: As described previously, only patients with outcomes recorded are retained 76 EVD PCR results are missing, resulting in 1721 (69%) patients being retained for the diagnosis model. 553 EVD+ patients did not have a confirmed final vital status (521 unknown, 28 transferred and 4 escaped) resulting in 1244 patients retained for the prognostic model.

Feature elimination and imputation

Clinical features. Grouping features according to their missingness, (See Table 3.2) the first group of symptoms have almost all observations known, the second group of symptoms are known only for 25 % and finally the last group is missing for 96% of them. Features of this latter group are dropped due to statistical irrelevance. For 1364/1797 patients, all of the group 2 symptoms are marked as "unknown". Using One hot encoding we group them all in a single Unknown feature with the value of 1 if all of the symptoms of group 2 are Unknown. There are 22 patients who don't have them all unknown but have only at least one value of group 2 symptoms that isn't unknown. We decide to remove them, 1797/2500 (72%) patients remain.

Redundancy. Some symptoms are redundant and very similar to each other. Vomiting and Nausea don't bring additional information to the feature Nausea / Vomiting. As they have 75% of Unknown values we therefore decided to not use these symptoms as features. We also drop the symptom Fatigue that is redundant and collinear with Asthenia.

Demographics. For the age 11/1797 patients had missing values which we impute with the median (30 years). Sex is missing from 4 patients.

Referral time is computed as the number of days between the start of the symptoms and their first admission (see Figure 3.4). The dates are converted to integers representing the day of the year. We have 1223 missing values of date of admission in Macenta, 203 missing values of date of admission in Guéckédou, 1224 missing values of referral time, 1223 missing values of date of contact with suspected case and 1616 missing values of date visit of funeral. They are imputed with the default 0 value. In the event of transfer (i.e. with two dates of admission in 2 different centers Macenta and Guéckédou) referral time is computed according to the first admission. The distribution of referral time is shown in the figure below.

Table 3.2 – Symptoms distribution

Group	Symptom	Yes	No	Unknown
Group 1	Fever	0.69	0.30	0.01
	Nausea / Vomiting	0.43	0.56	0.01
	Diarrhea	0.46	0.53	0.01
	Asthenia	0.67	0.32	0.01
	Hiccups	0.07	0.92	0.01
	Bleeding/Hemorrhage	0.12	0.87	0.01
	Headache	0.49	0.49	0.2
	Abdominal pain	0.36	0.62	0.02
	Arthralgia	0.40	0.58	0.2
	Anorexia	0.52	0.46	0.02
Group 2	Myalgia	0.12	0.13	0.75
	Fatigue	0.20	0.05	0.75
	Vomiting	0.12	0.13	0.75
	Nausea	0.13	0.12	0.75
	Breathing difficulties	0.04	0.21	0.75
	Dysphagia	0.05	0.20	0.75
	Chest pain	0.05	0.20	0.75
	Cough	0.05	0.20	0.75
	Sore throat	0.05	0.20	0.75
	Rashes	0.01	0.23	0.76
	Confusion / Disorientation	0.01	0.23	0.76
	Coma / Loss of consciousness	0.01	0.23	0.76
	Conjunctivitis (red eye)	0.06	0.17	0.76
	Jaundice (yellow connective tissues / gums / skin)	0.01	0.23	0.76
	Retro-orbital pain / photophobia	0.01	0.22	0.76
Group 3	Hematomas / petechiae / purpura	0	0.03	0.97
	Bleeding gums	0	0.0.2	0.97
	Bleeding from injection sites	0	0.03	0.97
	Epistaxis	0	0.03	0.97
	Meleana	0	0.03	0.97
	Hematemesis	0	0.03	0.97
	Dark vomiting	0	0.03	0.97
	Hemoptysis	0	0.03	0.97
	Vaginal bleeding	0	0.03	0.97
	Hematuria	0	0.03	0.97

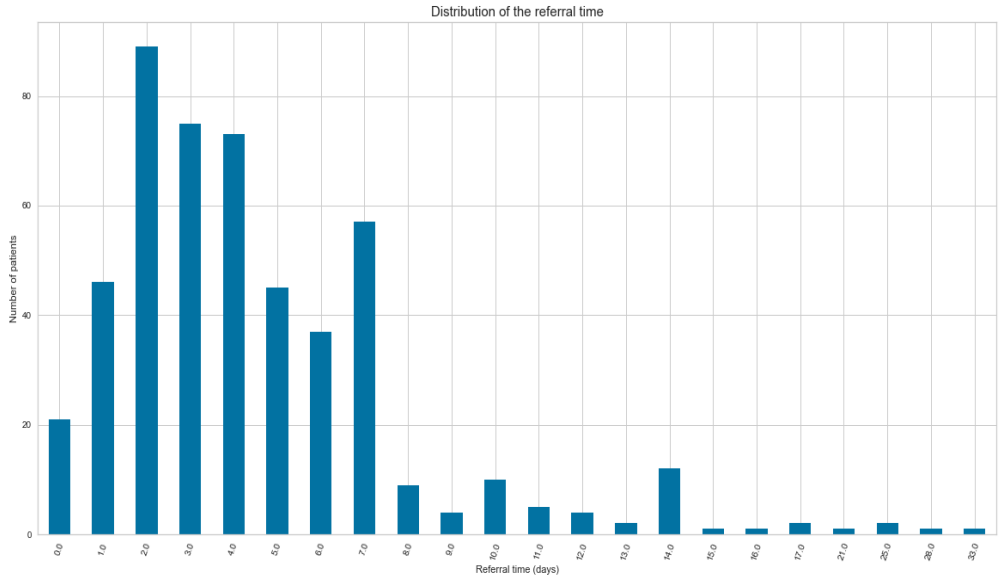


Figure 3.4 – Distribution of referral time

Chapter 4

Methods

4.1 Preprocessing and model selection

Train and test split We do not have a balanced number of samples for each class, this is a characteristic that we want to keep when we split the data set into train and test sets. We therefore use a stratified split with a test set proportion of 0.2. The cross validation set will be constructed using a 3-repeated stratified 5-fold split on the train set to tune the parameters.

Feature engineering We scale the numerical training data (age and dates) in the interval range [0,1] using a min max transformation. The Yes/No questions like patient currently hospitalized, link with a suspected case, visit of funeral and contact with body are one hot encoded to handle the missing values (Yes, No, Unknown).

Model selection In practice model selection is bit a more complicated than just select the 'best' algorithm, it's rather a process that we can decompose in three parts:

- Feature selection to obtain a set of predictive features
- Metric and family of algorithms selection
- Hyper parameter tuning for performance optimization

4.2 Feature selection

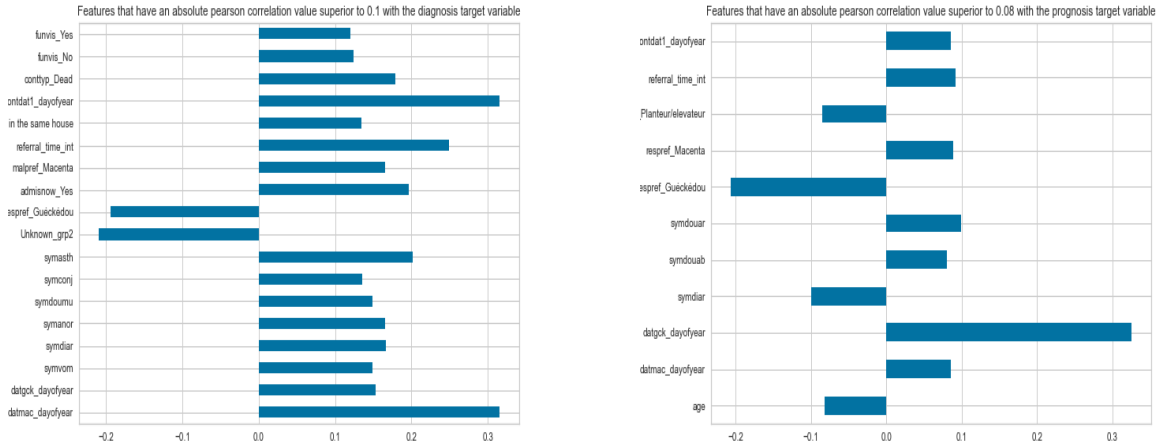
After encoding the data, we obtain 104 features but some of them are not statistically significant or can be sometimes very similar to other features. In order to avoid multicollinearity, overfitting and the curse of dimensionality, we proceed to a series of transformation to reduce the dimension of the data.

4.2.1 Variance threshold

We first apply a filter where we consider the variance of the features. To be useful, a features should discriminate between several samples so a feature with always the same value is not useful. All features with a variance below the threshold value 0.1 are removed (n=24).

4.2.2 Removing correlated features

Correlated features (see Figure 4.2) make the loss landscape ill-conditioned, by removing them we facilitate the convergence of the optimization algorithms and also avoid the problem of singular matrix. More importantly, our main objective is to have interpretable models and we lose this property when we deal with correlated features as we can't distinguish direct effect from indirect ones. If we take the example of regression analysis, we consider the coefficient of the dependent variable as the mean change needed from that variable to cause one unit change to the dependent variable while keeping all other dependent variables constant. When the features are correlated it's difficult to keep that assumption, as changing one variable will also influence the others. Consequently, the collinearity will not affect the precision of the model but rather the coefficients and their p-values in other words the model statistical power. We set a correlation threshold of 0.8 and unsurprisingly, we find that the features representing the residence prefecture and the prefecture where the patient got sick are almost equivalent, after removing also other correlated features, we end up excluding 9 features.



(a) Diagnosis data

(b) Prognosis data

Figure 4.1 – Univariate correlation with the dependent variable

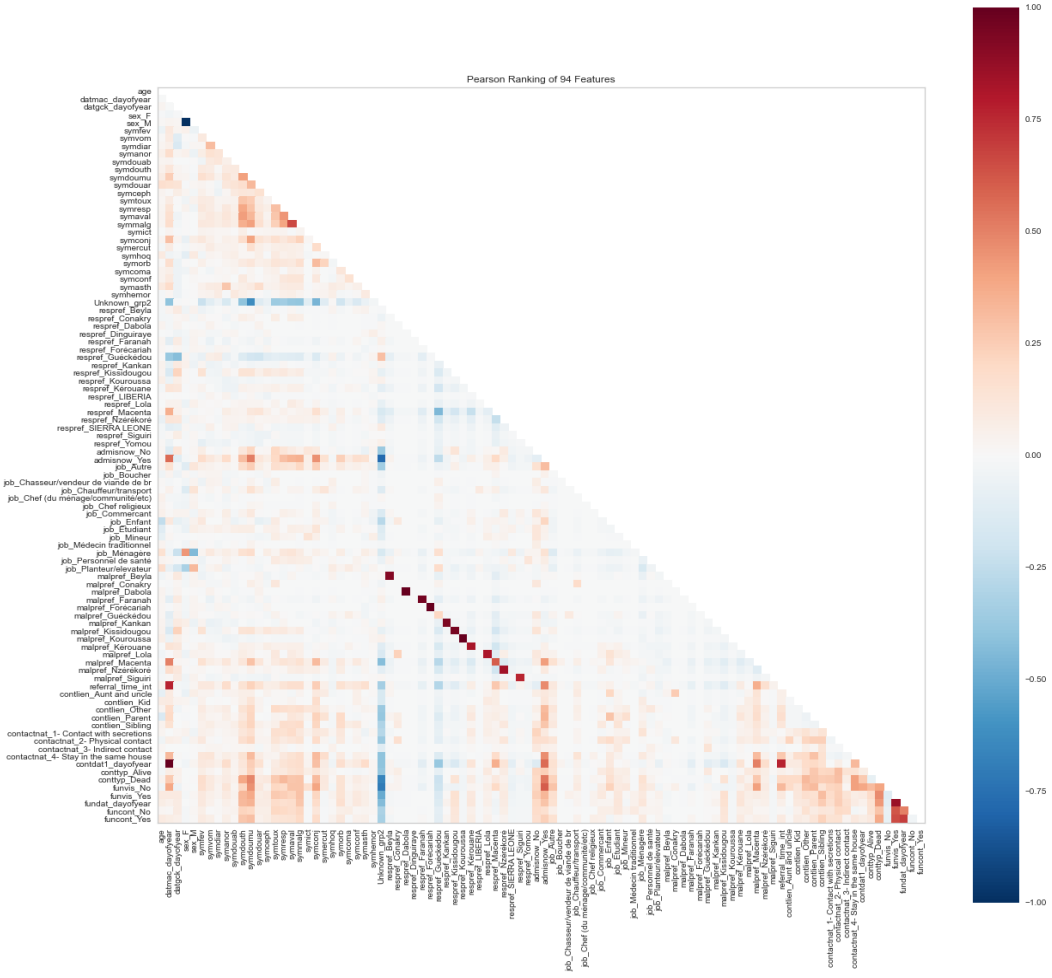


Figure 4.2 – Correlation between features

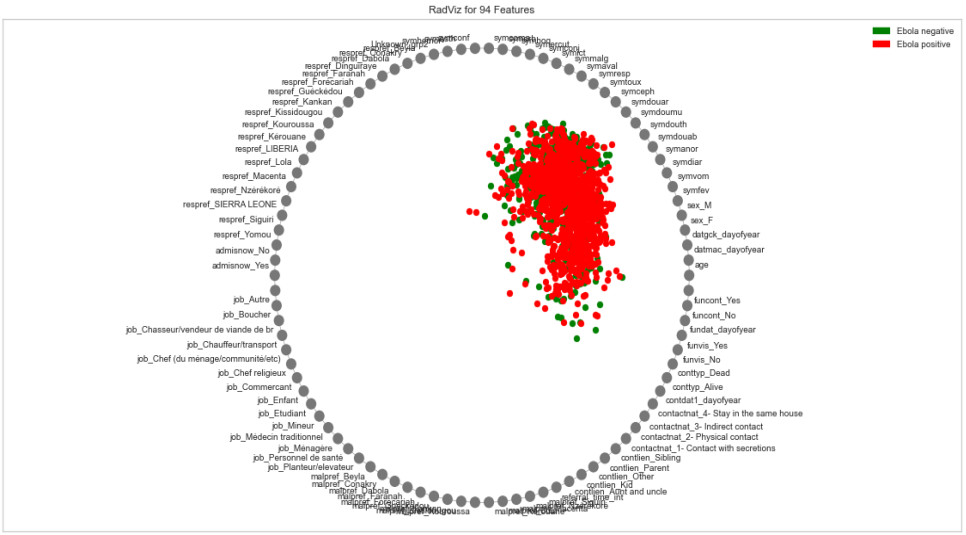


Figure 4.3 – RadViz of diagnosis data

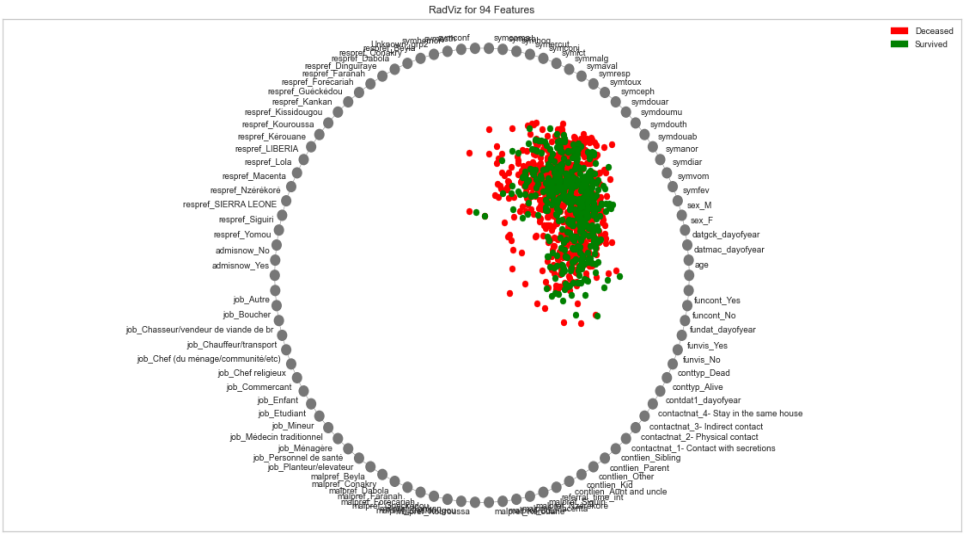


Figure 4.4 – Radviz of prognosis data

The sample distribution in the feature space (Figure 4.3 and Figure 4.4) highlights the difficulty of distinguishing the outcomes in this high dimension space. Nonetheless, we can observe

some differences and patterns that can allow us to differentiate between the two classes, a task that can be facilitated if we reduce the number of features. The simplest method would be to only keep the features that are correlated with the outcomes (see Figure 4.1) but this simple method would only capture linear interactions. In the following parts, we develop several other feature selection methods and present their benefits.

4.2.3 Recursive Feature Elimination (RFE)

The Recursive Feature Elimination (RFE) selects the best features by recursively removing the features with the least importance and constructing model with those that remain. For each model that has a feature importance metric we cross validate the number of features we keep. We can see an example of cross validation for logistic regression in Figure 4.5.

Table 4.1 – Results of the cross validation for the number of features selected by RFE

Task	Models	Number of features selected	AUC ROC CV
Diagnosis	Logistic	41	0.78
	Decision tree	19	0.70
Prognosis	Logistic	23	0.71
	Decision tree	1	0.60

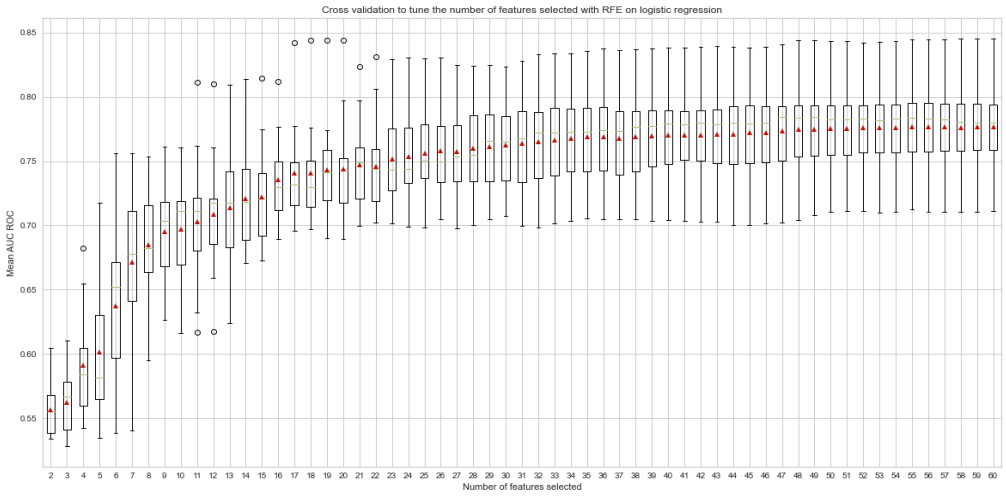


Figure 4.5 – Tuning of the number of features selected with RFE on logistic regression

4.2.4 Lasso selection and Analysis of Variance (ANOVA)

Linear regression with L1 norm regularization also known as Lasso regression promotes sparsity of features which we can leverage to select features. The features with the highest absolute coefficients are considered the most important. It results with 11 features with zero coefficients.

We want to find a score for each feature that determines how much it discriminates the dependent variable. Intuitively, to be a good discriminator a feature must have two properties. It should separate the means in order to have the two classes far from each other and each class should be grouped compactly together with a small within variance for each class. .The F-statistic captures these aspects with an increasing numerator when the classes are far a way and a decreasing denominator when the classes are compact.

For each model we use cross validation to find the best number of features to keep, ordered by their F-test score.

$$F = \frac{\text{between groups variance}}{\text{within group variance}} \tag{4.1}$$

In Table 4.2 for diagnosis, bagging trees performs well with very few features (17). The decision tree and KNN perform relatively well also compared to other models as they use only 1/5 of their features.

Diagnosis The features that the decision tree selects for diagnosis:

- Chronological features: Date of admission in Macenta center
- Geographical features: Residence prefecture is Gueckedou, Residence prefecture is Macenta
- Symptoms: Diarrhoea, Myalgia, Conjunctivitis and Weakness
- The feature that informs if the group 2 symptoms are all unknown
- Patient is currently hospitalized
- Referral time
- Contact tracing information with suspected cases: Date of contact and if she/he is dead

KNN uses the same features as above along with 3 additional ones: anorexia, date of admission in Gueckedou center and if the patient attended a funeral.

Prognosis For prognosis, with an AUC of 0.65 and using a subsset of 11 features SVM performs as well as using all the features. The features used are:

- Chronological features: Date of admission in Macenta center and Gueckedou center
- Geographical features: Residence prefecture is Gueckedou, residence prefecture is Macenta and whether the prefecture where the patient got sick is Beyla
- Symptoms: Diarrhoea, abdominal pain, joint pain and confusion
- Patient is a farmer
- Referral time

We remark that similarly as for Recursive Feature Elimination the decision tree only uses 1 feature for prognosis it's the day of admission in Gueckedou center. This points to the fact that the model is just predicting daily prevalence rather than prognosis.

Table 4.2 – Cross validation for number of features selected ordered by F-test score

Task	Models	Number of features selected	Best mean AUC ROC CV
Diagnosis	SVM	61	0.79
	KNN	15	0.76
	Logistic	61	0.78
	Decision tree	12	0.75
	Bagging trees	17	0.80
	Extra trees	58	0.78
	Random Forest	59	0.81
Prognosis	SVM	11	0.66
	KNN	20	0.66
	Logistic	22	0.70
	Decision tree	1	0.60
	Bagging trees	30	0.67
	Extra trees	37	0.64
	Random Forest	37	0.68

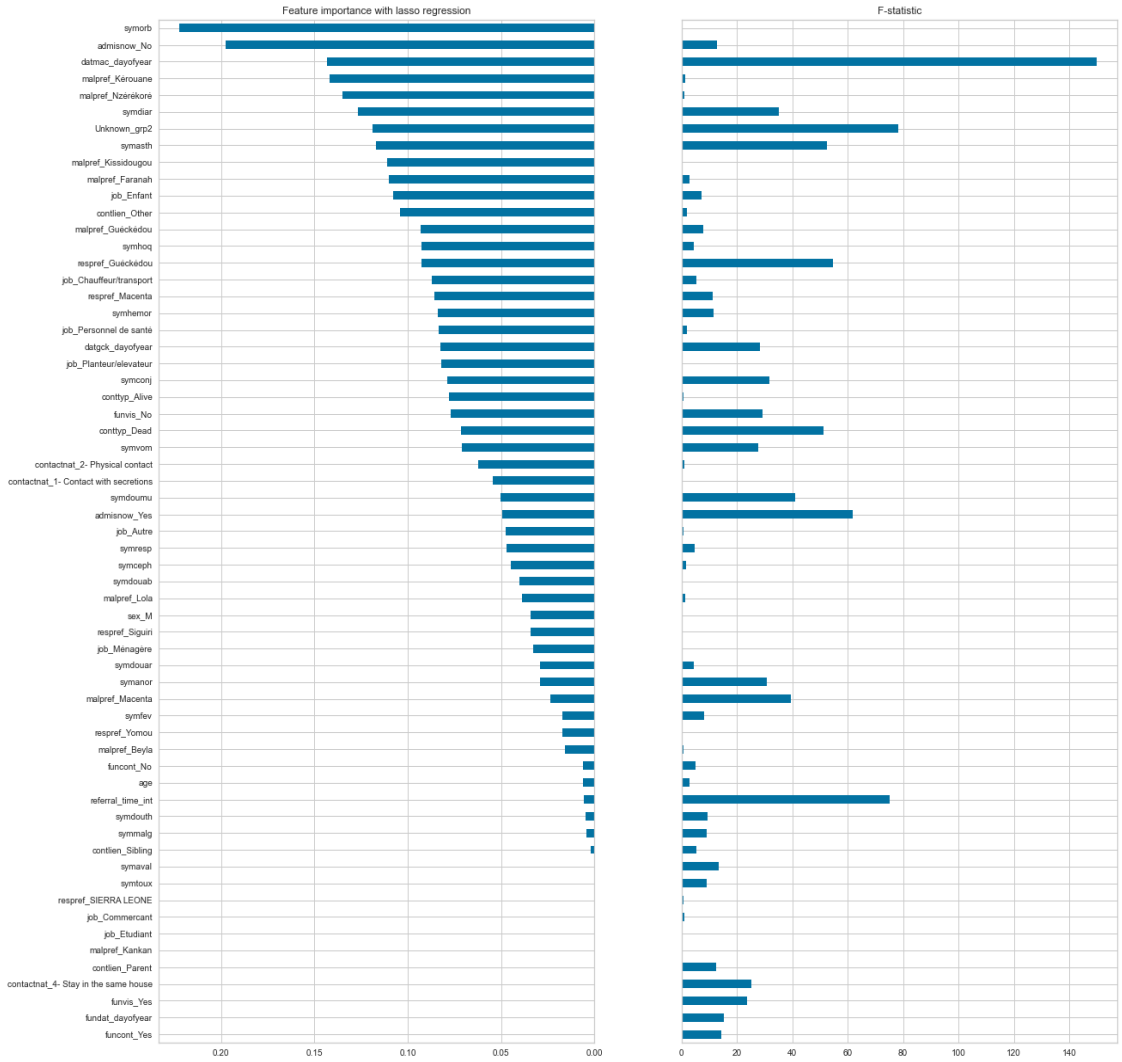


Figure 4.6 – The coefficients of Lasso regression and the F-statistics

Figure 4.6 shows clearly that the features are not ranked in the same way between Lasso elimination and F-test. Indeed, light sensitivity for example has the biggest coefficient in lasso elimination whereas its F-score is null. This is can be interpreted by the fact that the two methods don't prioritize the same properties when they select features.

4.2.5 Random Forest Feature Importance and Permutation importance

In decision tree, every node states a rule to split the data with a respect to a feature. The objective is to maximize purity of the leaves or equivalently to minimize the Gini impurity. When training a tree we can compute how much each feature contributes to decreasing the weighted impurity.

The permutation feature importance is the decrease in a model score when a single feature value is randomly shuffled. The decrease of the model score shows how much the the feature participates in the predictions. The advantage of that technique is that it's model agnostic as it only affects the data.

Figure 4.7 – Feature importance with permutation on Random Forest

The permutation importance plot shows that permuting a feature drops the AUC ROC by at most 0.025, which would suggest that none of the features are important.

When features are collinear, permuting one feature will have a small effect on the model's performance because it can get the same information from a correlated feature. One way to handle multicollinear features is by performing hierarchical clustering on the Spearman's correlations, picking a threshold, and keeping a single feature from each cluster.

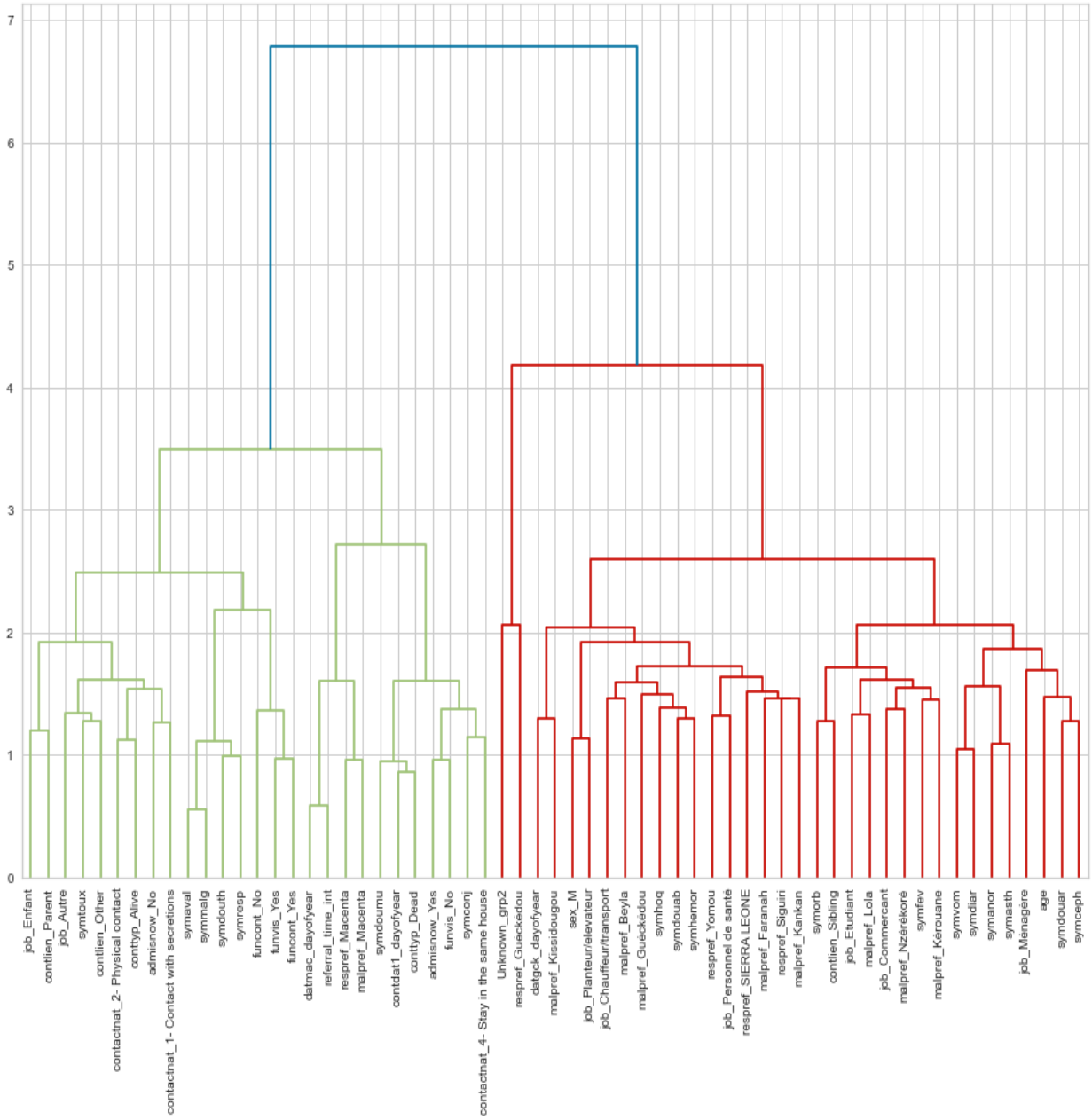


Figure 4.8 – Hierrarchical Agglomerative Clustering (HAC) with Ward's linkage for diagnosis

We compute a Hierarchical Agglomerative Clustering (HAC) with Ward's linkage to approximate the partition minimizing the error sum of squares. The distance between two clusters is how much the sum of squares will increase when we merge them. With hierarchical clustering having initially every point is in its own cluster, the sum of squares starts out at zero and then grows as we merge clusters. Ward's method keeps this growth as small as possible.

Finally, we manually pick a threshold by visual inspection of the dendrogram to group our features into clusters and choose a feature from each cluster to keep. If we set the threshold to 1 for diagnosis, eight features are removed and 53 features remain while for prognosis 7 features are removed and 55 features remain.

4.3 Metrics (optional)

We advocate here the choice to evaluate the models and tune their parameters using the AUC ROC metric. A reader familiar with machine learning can skip this part, for the others: bear with me for few minutes.

Let's start by defining a "ground" naive choice that we should aim to improve. As we have seen before, we have an unbalanced set, in any situation with unbalanced sets, a naive estimator that assumes the majority class to be a solution, will give us a very high accuracy.

Before we continue, let's discuss the few most important metrics to evaluate our model, and how we can present the results:

- **Accuracy**: Measures how well our model predicts all the classes, regardless of balance. It is the ratio of "correctly predicted" results, versus the entire sample.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.2}$$

  If we are agnostic about our categories, aiming for the highest accuracy possible is generally good enough. However, in medicine, where preliminary tests might indicate the presence or absence of a disease. If the prevalence is 1%, that means that 99% of accuracy will be predicted by just discarding the disease. We are however not interested in correctly predicting 99% of the cases, but rather on finding as many as possible of the remaining 1%. In other words, it's better to be wrongly diagnosed with a rare disease for further testing, rather than be wrongly diagnosed as healthy when you actually have the disease.

- **Precision**: Measures the fraction of actually positive cases among those predicted to be positive.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.3}$$

  Precision is important when false positives are more costly than false negatives, not our

case. A trivial way to have perfect precision is to make one single positive prediction and ensure it is correct (precision = 1/1 = 100%). This would not be very useful since the classifier would ignore all but one positive instance. So precision is typically used along with another metric, recall.

- **Recall**: Measures the fraction of actually positive cases found from all positive cases.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.4}$$

  Recall is important when we aim to find the positive cases, like in this study.

- **F1 score**: It's the harmonic mean of precision and recall, usually a good metric of the balance between the two metrics.

$$\text{F1} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{4.5}$$

  The F1 score favors classifiers that have similar precision and recall. Unfortunately, we can't have it both ways: increasing precision reduces recall, and vice versa.

- **ROC AUC**: The receiver operating characteristic (ROC) curve is another common tool used with binary classifiers. It plots recall (true positive rate) against the false positive rate (FPR), that is the ratio of negative instances that are incorrectly classified as positive. FPR is equal to one minus the true negative rate (TNR), which is the ratio of negative instances that are correctly classified as negative. The TNR is also called specificity. Hence the ROC curve plots recall versus 1 – specificity.

  ROC is a probability curve and the area under the curve (AUC) is a measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, better the model is at distinguishing between patients with disease and no disease.

The unbalanced data set makes it difficult to optimize for the recall or other simple metrics, so we decide to evaluate and tune the parameters over ROC AUC metric to obtain robust models.

Sensitivity which measures the portion of positive people that are correctly classified is the most important metric. However, specificity which measures the portion of negatives that are correctly classified becomes more important when the patients arrive at the hospital in order to optimize the limited resources.

Chapter 5

Results

We develop here the different models that we tried, their hyper parameter tuning and their interpretabilty.

5.1 Logistic regression

Logistic regression is well suited for discovering links between features and outcomes.

We consider a single input observation represented by a vector of features $x^{(i)}$ = [$x_1^{(i)}$, $x_2^{(i)}$,..., $x_d^{(i)}$]. The classifier can output y = 1 predicting that the patient is EVD positive or that they will survive (or the opposite by outputting 0).

For a given patient $x^{(i)}$, we aim to find the probability of belonging to the positive class :

$$P(y^{(i)} = 1) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_d x_d^{(i)}))} \tag{5.1}$$

The interpretation of the weights in logistic regression is more cumbersome than in liner regression the outcome being in logistic regression a probability. With the logistic function the weights do not influence the probability linearly any longer. Therefore we need to reformulate the equation for the interpretation so that only the linear term is on the right side of the formula.

$$log\left(\frac{P(y=1)}{1-P(y=1)}\right) = log\left(\frac{P(y=1)}{P(y=0)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d \tag{5.2}$$

We have shown that the logistic regression model is a linear model for the log odds. We know ask ourselves how the prediction changes when one of the features x_j is changed by 1 unit.

$$odds = \frac{P(y=1)}{1-P(y=1)} = exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\right) \tag{5.3}$$

$$\begin{aligned} \frac{odds_{x_j+1}}{odds} &= \frac{exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_j(x_j+1) + \ldots + \beta_p x_p\right)}{exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_j x_j + \ldots + \beta_p x_p\right)} \\ &= exp\left(\beta_j(x_j+1) - \beta_j x_j\right) \\ &= exp\left(\beta_j\right) \end{aligned} \tag{5.4}$$

We conclude that a change in a feature by one unit changes the odds ratio by a factor of $exp\left(\beta_j\right)$

We tune the regularization by modifying its norm and value. We can see an example for the L2 norm in Figure 5.1

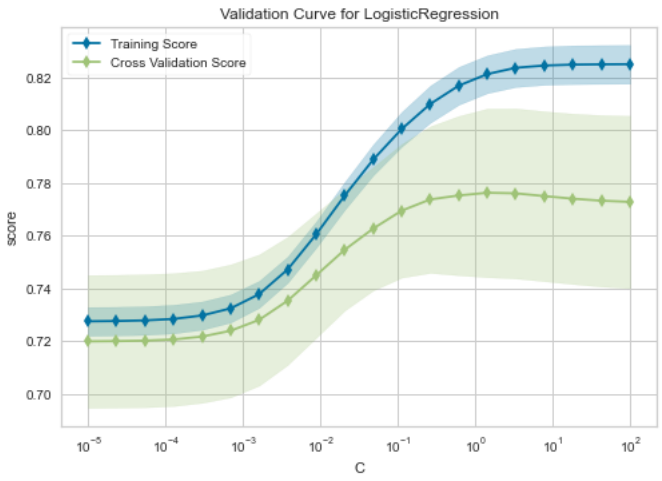


Figure 5.1 – Hyperparameter tuning of the regularization term value

We don't have significant differences between the 4 feature selection methods for diagnosis and prognosis (see Table 5.1) For **diagnosis**, RFE required the lowest number of features (n=41) features to achieve a similar performance (AUC of 0.8) to other models which required at least 10 more features. For **prognosis**, Lasso required the least number of features (n=5) to reach a similar performance than the other models requiring over 15 features more.

Table 5.1 – Results for logistic regression

Task	Feature selection	Number of features	Training set			Test set		
			Accuracy	AUC	F1-score	Accuracy	AUC	F1-score
Diagnosis		61	0.74	0.82	0.81	0.70	0.77	0.79
	RFE	41	0.74	0.81	0.80	0.72	0.77	0.80
	F-test	61	0.74	0.82	0.81	0.70	0.77	0.79
	Lasso	52	0.75	0.81	0.81	0.70	0.76	0.79
	HAC	53	0.74	0.81	0.81	0.71	0.77	0.79
Prognosis		62	0.69	0.74	0.45	0.63	0.67	0.34
	RFE	23	0.69	0.74	0.47	0.64	0.64	0.40
	F-test	22	0.69	0.74	0.5	0.65	0.68	0.41
	Lasso	5	0.67	0.70	0.40	0.63	0.66	0.32
	HAC	56	0.68	0.74	0.45	0.62	0.65	0.35

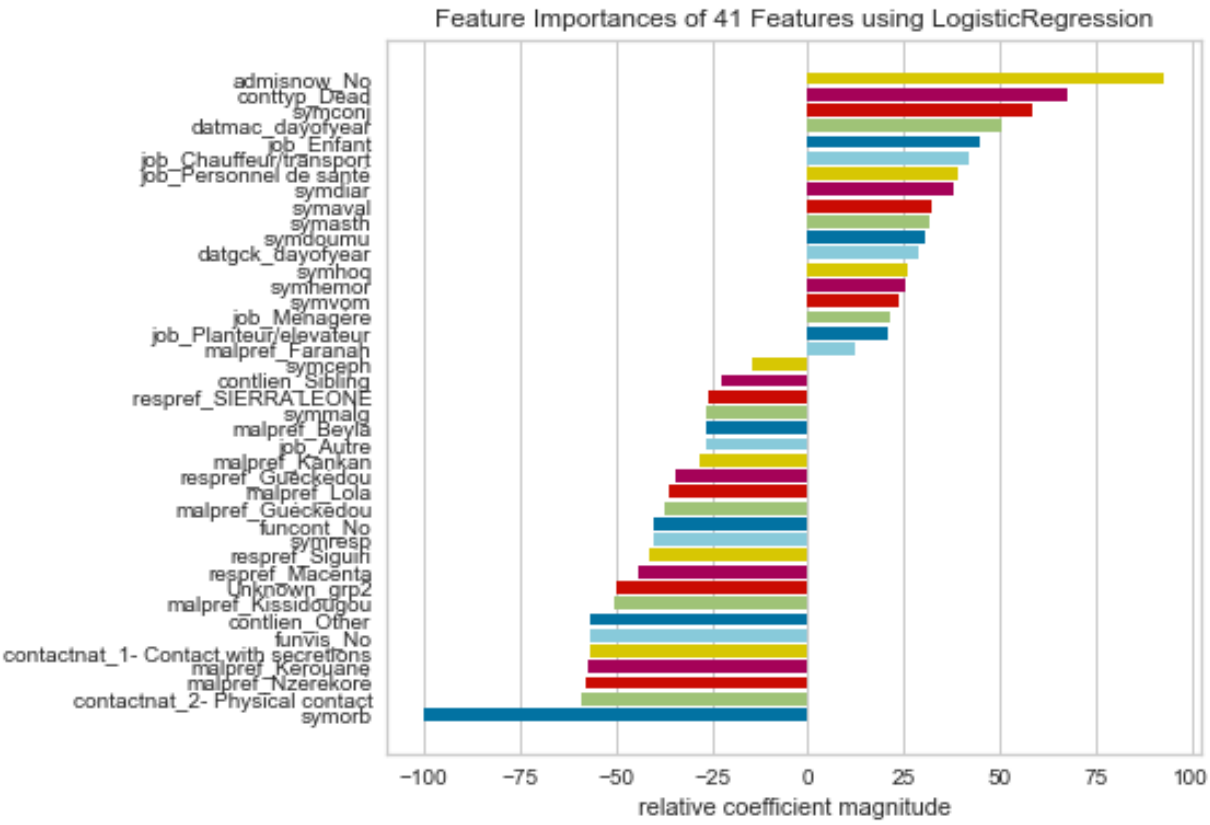


Figure 5.2 – Logistic regression after RFE for diagnosis

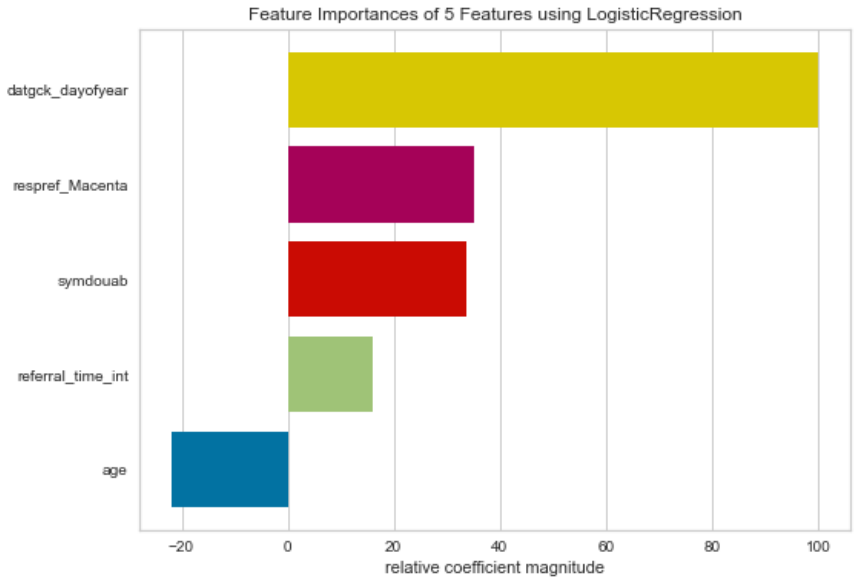


Figure 5.3 – Logistic regression after lasso elimination for prognosis with 5 features : day of admission in Gueckedou, residency in Macenta, abdominal pain, referral time and age

5.2 SVM

For classification tasks, logistic regression doesn't consider the proximity of the points to the decision boundaries. Intuitively, if a point is further a way from the decision boundary we can be more confident in our prediction. Consequently, the optimal decision boundary is the boundary that maximizes its margin with all data points. Support Vector Machines (SVM) find that decision boundary.

A standard SVM aims to find a decision boundary with a maximal margin that separates all positive and negative examples. However, this can typically lead to over fitting and high variance especially with noisy data. We can make the model more robust with a "soft margin" allowing some examples to be miss classified and to not be considered for the margin. This trade off between training error and robustness is controlled by the regularisation parameter C that we tune.

While the feature selection based on F-test score don't remove any features after cross validation (see Table 5.2). HAC and Lasso with 10 less features perform the same with an AUC of 0.81. Furthermore, SVM outperforms logistic regression for diagnosis and give equivalent results for prognosis.

Table 5.2 – Results for SVM

Task	Feature selection	Number of features	Training set			Test set		
			Accuracy	AUC	F1-score	Accuracy	AUC	F1-score
Diagnosis		61	0.80	0.88	0.85	0.73	0.81	0.79
	F-test	61	0.80	0.88	0.85	0.73	0.80	0.80
	Lasso	52	0.79	0.87	0.85	0.73	0.81	0.81
	HAC	53	0.81	0.89	0.86	0.72	0.81	0.80
Prognosis		62	0.71	0.78	0.40	0.64	0.67	0.27
	F-test	11	0.67	0.71	0.29	0.61	0.62	0.13
	Lasso	5	0.67	0.68	0.18	0.64	0.65	0.12
	HAC	56	0.70	0.77	0.40	0.63	0.64	0.28

5.3 KNN

The K Nearest Neighbor (KNN) classifier is a memory based algorithm and therefore it doesn't require any training. Given a patient, the algorithm finds the k nearest nearest patients in the feature space and applies a majority vote to output the patient class. We tune how we define distance between 2 points (Manhattan distance or Euclidean distance) and the number k of neighbors considered (see Figure 5.4).

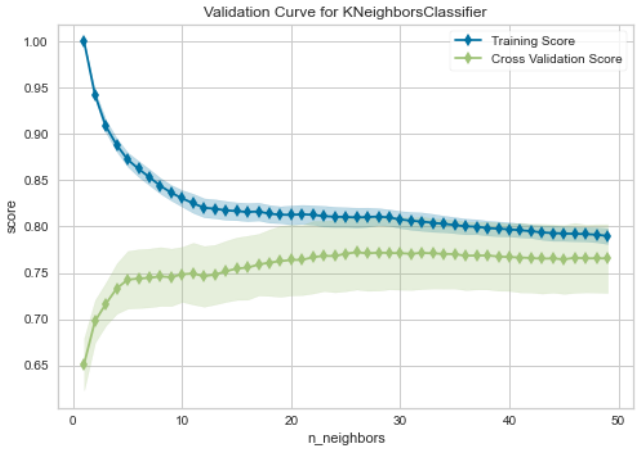


Figure 5.4 – Cross validation for the number k of neighbors

For diagnosis with the selection based on F-test score after cross validation we reduce the number of features to 15 and obtain an AUC of 0.82 slightly superior to SVM. For prognosis the results are equivalent than the other models (see Table 5.4).

Table 5.3 – Results for knn

Task	Feature selection	Number of features	Training set			Test set		
			Accuracy	AUC	F1-score	Accuracy	AUC	F1-score
Diagnosis		61	0.73	0.81	0.81	0.70	0.78	0.79
	F-test	15	0.78	0.86	0.85	0.73	0.82	0.81
	Lasso	52	0.73	0.81	0.81	0.71	0.78	0.79
	HAC	53	0.84	0.82	0.81	0.72	0.79	0.79
Prognosis		62	0.71	0.76	0.52	0.65	0.68	0.46
	F-test	20	0.69	0.75	0.48	0.65	0.66	0.40
	Lasso	5	0.72	0.79	0.52	0.64	0.68	0.35
	HAC	56	0.70	0.75	0.53	0.64	0.67	0.42

5.4 Decision tree

Logistic regression models fail in situations where the relationship between features and outcome is nonlinear or where features interact with each other. Tree models iteratively partition the data on certain values of the features.

Several algorithms are used to create trees and they differ by structure. The classification and regression trees (CART) algorithm is probably the most popular algorithm for tree induction.

To predict the outcome in each leaf node, the average outcome of the training data in this node is used.

$$\hat{y} = \hat{f}(x) = \sum_{m=1}^{M} c_m I\{x \in R_m\} \tag{5.5}$$

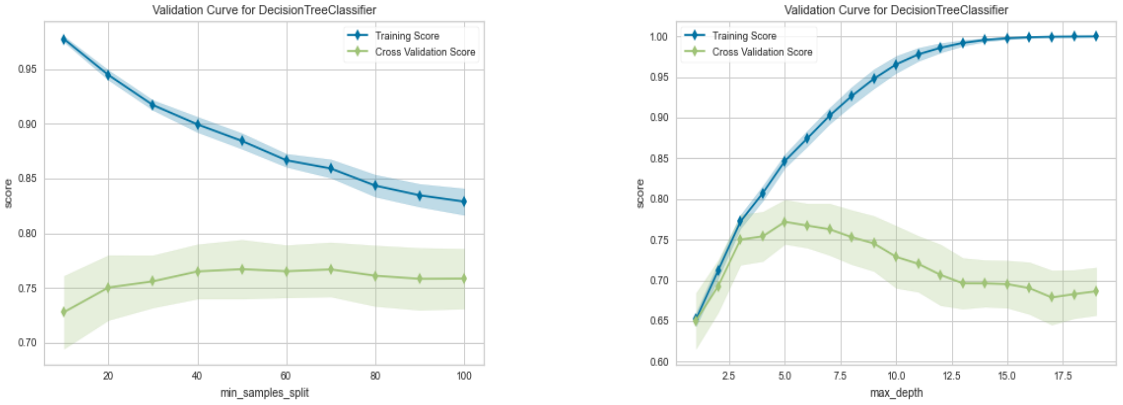
$I\{x \in R_m\}$ is the identity function that returns 1 if x is in the subset R_m and 0 otherwise.

How does CART construct the tree? (in other words, how does it select the features on which it splits the data?) CART takes a feature and determines which cut-off point minimizes the Gini index of the class distribution of y. The Gini index tells us how "impure" a node is, if all classes have the same frequency, the node is maximally impure, if only one class is present, it is "pure". The Gini index is minimized when the data points in the nodes have very similar values for y. The best cut-off point makes the two resulting subsets as different as possible with respect to the target outcome.

The main drawback of decision trees is that they have many parameters to tune and tend to over fit the training set.

The parameters on which we focused are:

- Maximum depth: The deeper the tree, the more splits it has and it captures more information about the data.
- Minimum number of samples to split: It represents the minimum number of samples required to split an internal node. When we increase this parameter, the tree becomes more constrained as it has to consider more samples at each node.
- Maximum number of features: It represents the number of features (randomly selected) to consider when looking for the best split.



(a) Cross validation of the minimum number of split

(b) Cross validation of the max depth

Figure 5.5 – Validation curve of decision tree parameters

Table 5.4 – Results for decision tree

Task	Feature selection	Number of features	Training set			Test set		
			Accuracy	AUC	F1-score	Accuracy	AUC	F1-score
Diagnosis		61	0.77	0.85	0.82	0.74	0.80	0.80
	RFE	19	0.78	0..87	0.84	0.75	0.80	0.82
	F-test	12	0.71	0.79	0.79	0.75	0.77	0.82
	Lasso	52	0.79	0.88	0.84	0.71	0.78	0.78
	HAC	53	0.77	0.86	0.82	0.74	0.80	0.80
Prognosis		62	0.70	0.76	0.50	0.65	0.67	0.40
	RFE F-test	1	0.69	0.75	0.44	0.58	0.66	0.27
	Lasso	5	0.70	0.77	0.59	0.64	0.70	0.50
	HAC	56	0.73	0.80	0.55	0.62	0.63	0.33

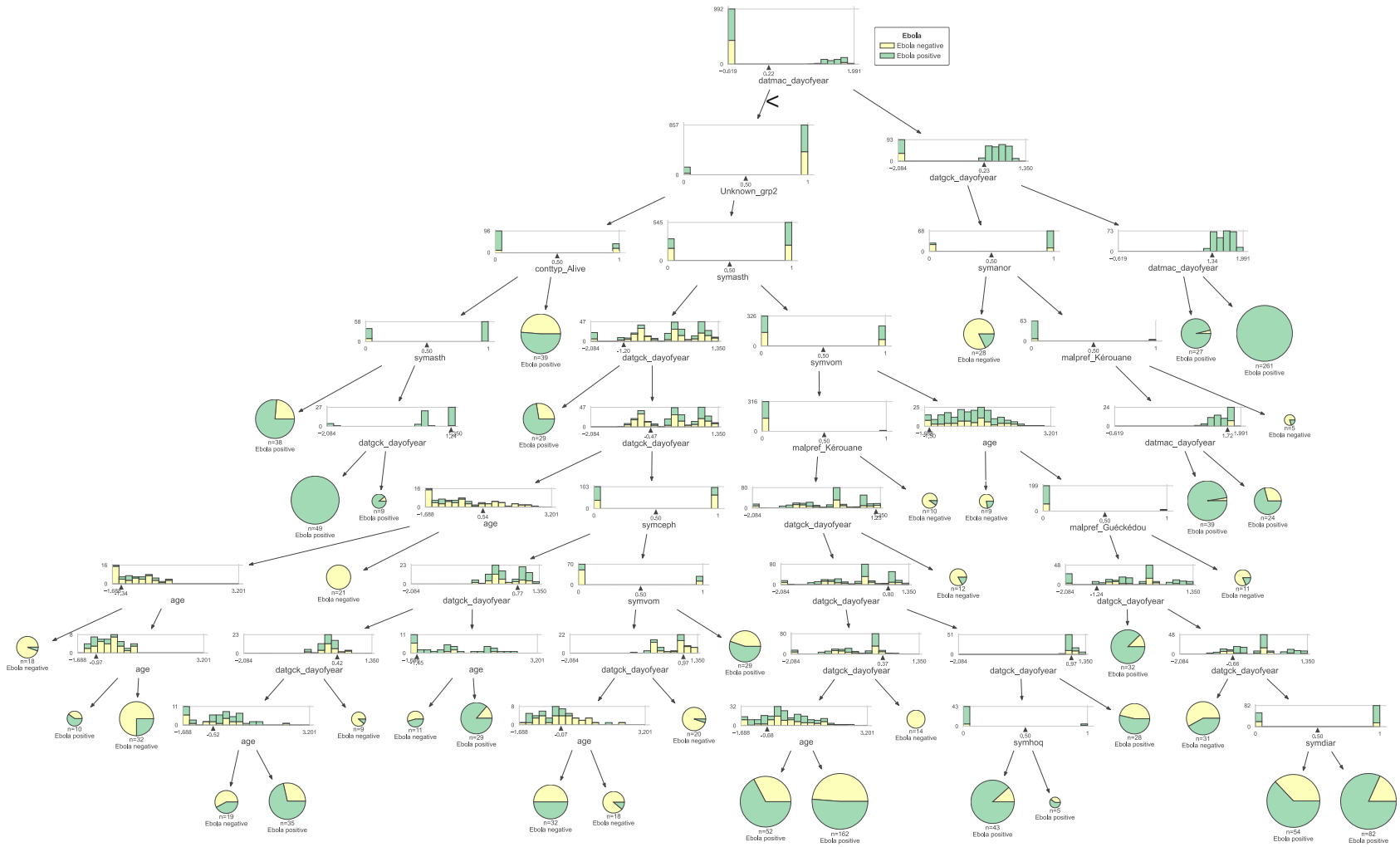


Figure 5.6 – Decision tree

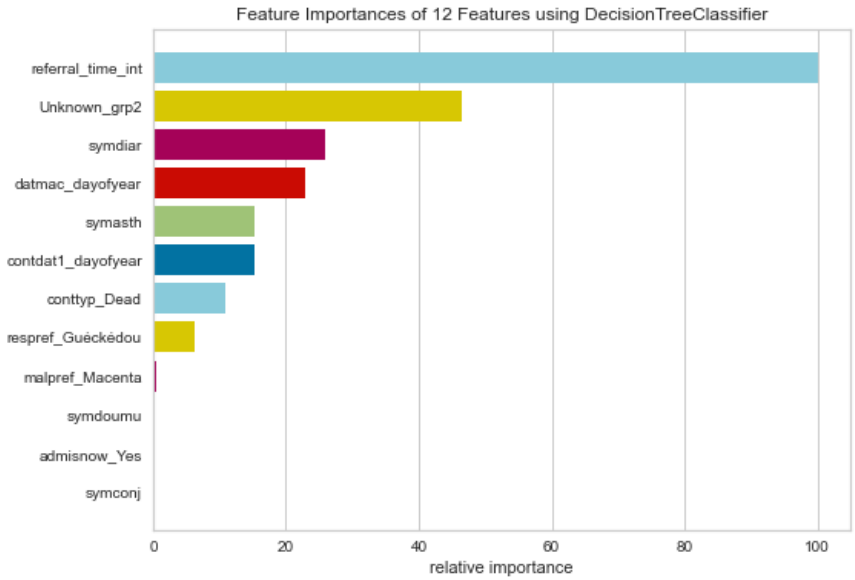


Figure 5.7 – Decision tree after F-test selection for diagnosis

5.5 Ensemble Methods : Wisdom of the crowd

To answer complex questions, it's usually better to have several expert opinions. That idea gives rise to ensemble learning where a model aggregates the predictions of many predictors to make a final prediction.

5.5.1 Bagging

Bagging (Bootstrap Aggregating) classifier is an ensemble method where each weak learner is trained with a random subset of the data sampled with replacement (bootstrapping). It then combines the weak learner's predictions by outputting their average.

With 90 decision trees as weak learners, bagging slightly outperforms the other methods. After F-test selection for diagnosis we obtain on the training set an Accuracy of **0.79**, an AUC of **0.88** and F1 score of **0.85** while on the test set we get an Accuracy of **0.76**, an AUC of **0.84** and F1 score of **0.83**.

For prognosis we also we get a small improvement with on the training set an Accuracy of **0.78**, an AUC of **0.84** and F1 score of **0.63** and on the test set an Accuracy of **0.6**, an AUC of **0.72** and F1 score of **0.44**.

5.5.2 Extra trees and Random Forest

Decision trees suffer from being high-variance estimators, the addition of a small number of extra training observations can dramatically alter the prediction performance of a learned tree, despite the training data not greatly changing. By training weak learners on different subsets of data the ensemble methods Extra Trees and Random Forest allow us to avoid over fitting with the assumption that the aggregation step will eliminate the variance error.

Extra Trees and Random Forest are both composed of a large number of decision trees, where the final decision is obtained by majority vote. The main two differences are the following:

- Random forest uses bootstrap replicas, for each tree it sub samples the input data with replacement, whereas Extra Trees use the whole original sample which may increase variance because bootstrapping makes it more diversified.
- Random Forest chooses the optimum split while Extra Trees selects it randomly. However, once the split points are fixed, the two algorithms choose the best one among the entire subset of features. Therefore, Extra Trees adds randomization but still has optimization.

These differences motivate the reduction of both bias and variance. On one hand, using the whole original sample instead of a bootstrap replica will reduce bias but increase variance. On the other hand, choosing randomly the split point of each node will reduce variance but increase bias.

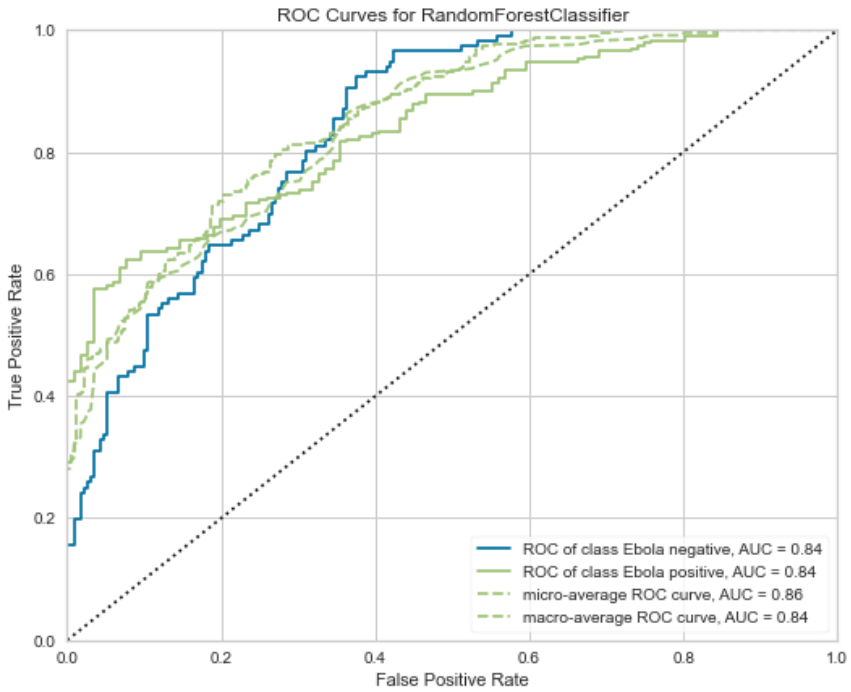


Figure 5.8 – ROC curve for Random Forest diagnosis

Extra trees and Random Forest perform similarly to bagging classifiers with on the test set an **AUC of 0.84** for diagnosis (see Figure 5.8) and an **AUC of 0.73** for prognosis. We can observe a full analysis of Random Forest predictions on 5.9

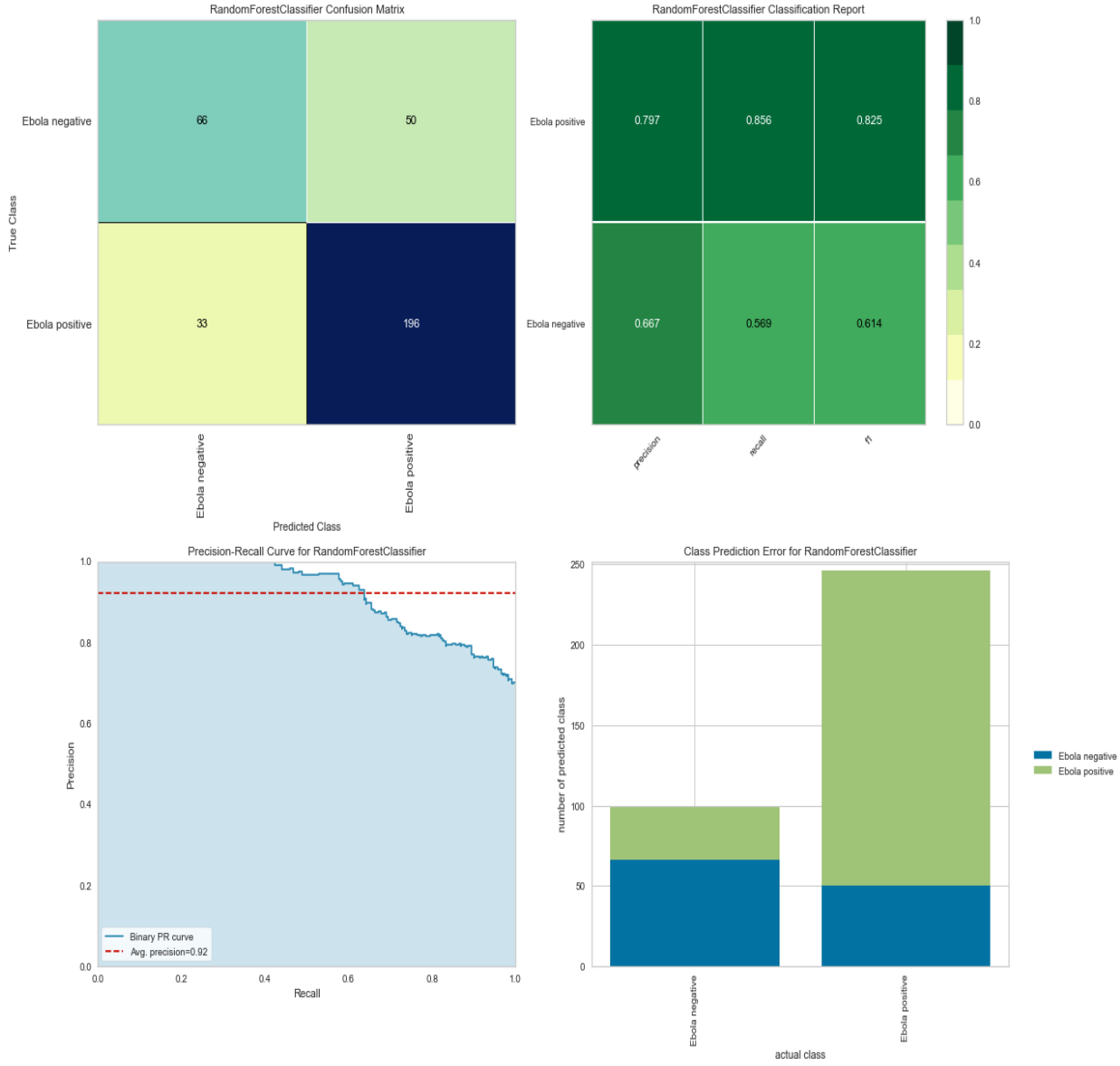


Figure 5.9 – Results on the test set for Random Forest diagnosis

5.6 Comparison with another study

The exact same pipeline has been applied to another study (EIXUZQ) with a smaller cohort size of 870 patients where very similar information and symptoms have been collected (see Figure 5.10)

Diagnosis The best results are obtained with the Extra tree ensemble method without feature selection, where we obtained on the test set an **AUC of 0.77**. This is 0.07 lower than in the first data set and used 47 features (compared to 58 features above). Interestingly the features selected were similar as shown in Figure 5.12.

Prognosis Here, lasso elimination retained just 3 features and achieved an **AUC of 0.85**. We can explain the difference for the prognosis model by the fact that for the first study the cycle threshold (CT) value of the EVD test which is the most important feature (see Figure 5.11) is not present which makes it harder to predict the patient outcome. The CT value is the viral load quantified by PCR and has been well established as a strong prognostic indicator.

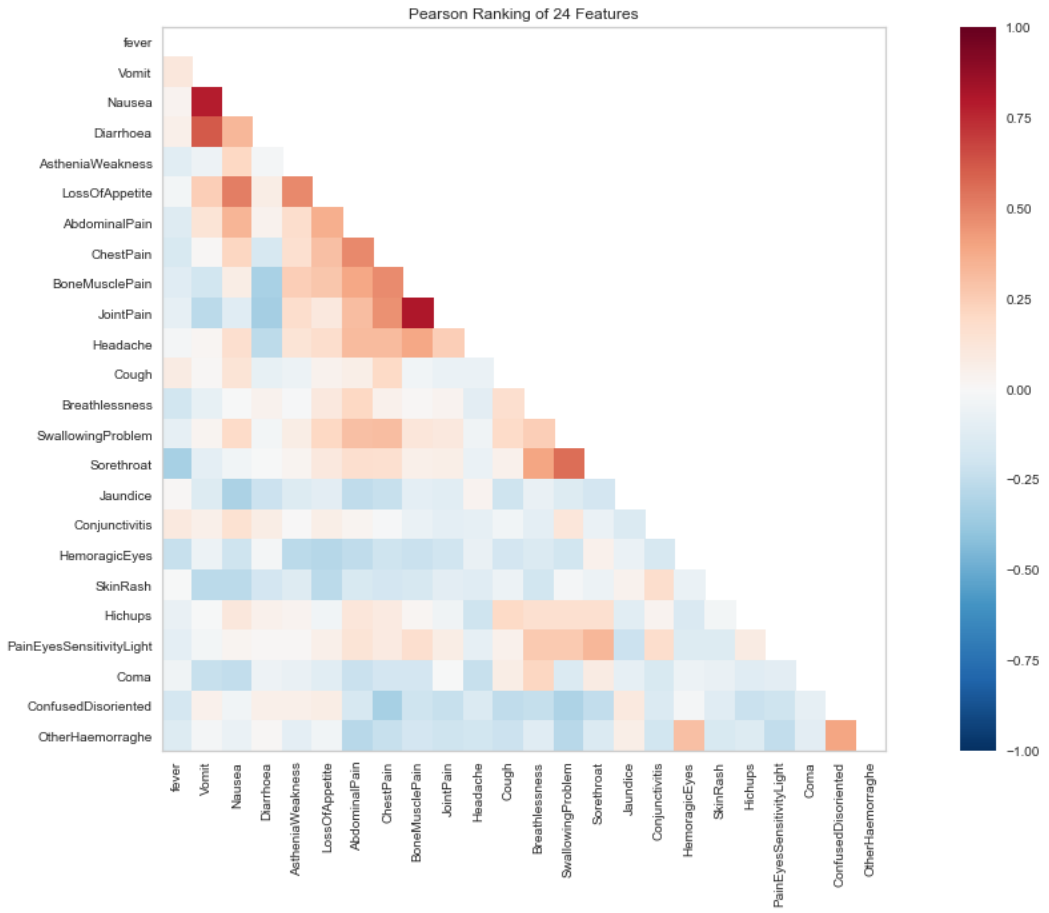


Figure 5.10 – Correlation between the symptoms

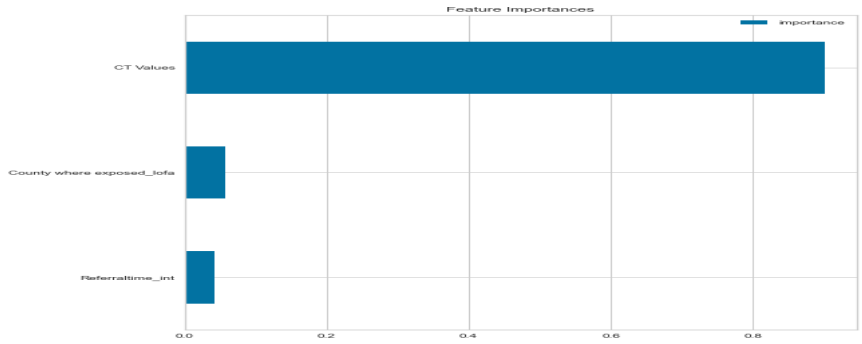


Figure 5.11 – Feature importance with Extra trees for prognosis

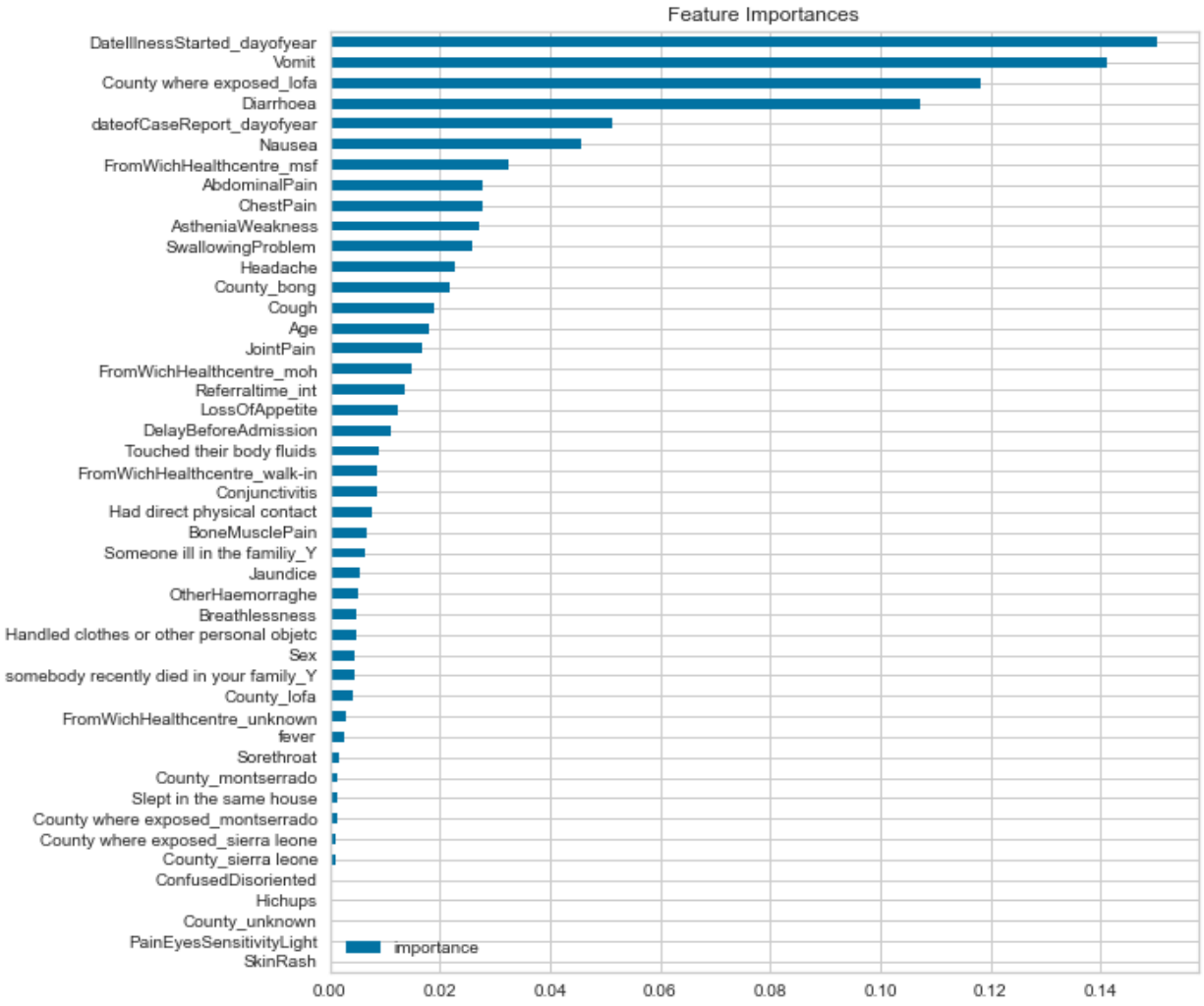


Figure 5.12 – Feature importance with Extra trees for diagnosis

Chapter 6

Conclusion

Taken all together Machine Learning in general, has a lot of potential to predict EVD diagnosis and severity and thus guide evidence-based triage decisions at minimal cost.

We emphasize the importance of exploring different feature selection methods and models and prioritize interpretable methods that allow the user to understand the rationale of collecting the required information and the importance of estimating it correctly. It's hard to compare our results with existing classical statistical (non-machine learning) methods as their evaluation is not calculated in the same way (boostrapping vs an independent test set). With an AUC above 0.8 for diagnosis and prognosis on test sets, using a reasonable number of features, we have obtained scores that make the models a promising and feasible addition to improve triage in EVD.

It has been shown that more data improves predictive performance and robustness. Therefore it is important to join the different studies. While there is a significant overlap across the different studies, when we compared them we have encountered a lot of heterogeneity and systematic differences that need to be aligned. This project lays the groundwork for more general models trained on several studies leveraging distributed and federated learning.

Bibliography

[1] Hartley M-A, Young A, Tran A-M, OkoniWilliams HH, Suma M, Mancuso B, et al. (2017) *Predicting Ebola Severity: A Clinical Prioritization Score for Ebola Virus Disease.* PLoS Negl Trop Dis 11(2): e0005265. doi:10.1371/journal. pntd.0005265

[2] Hartley MA, Young A, Tran AM, et al. *Predicting Ebola infection: A malaria-sensitive triage score for Ebola virus disease.* PLoS Negl Trop Dis. 2017;11(2):e0005356. Published 2017 Feb 23. doi:10.1371/journal.pntd.0005356

[3] Joel G. Breman et al. *Discovery and Description of Ebola Zaire Virus in 1976 and Relevance to the West African Epidemic During 2013–2016.* The Journal of Infectious Diseases 214 (Suppl 3 Oct. 15, 2016).

[4] Weyer J, Grobbelaar A, Blumberg L. *Ebola virus disease: history, epidemiology and outbreaks. Current infectious disease reports.* 2015 May; 17(5):480. doi: 10.1007/s11908-015-0480-y PMID: 25896751

[5] Green A. *DR Congo Ebola virus treatment centres attacked.* Lancet. 2019 Mar 16;393(10176):1088. doi: 10.1016/S0140-6736(19)30576-8. Epub 2019 Mar 14. PMID: 30957744.

[6] Jadav SS, Kumar A, Ahsan MJ, Jayaprakash V. *Ebola virus: current and future perspectives. Infect Disord Drug Targets.* 2015;15(1):20-31. doi: 10.2174/1871526515666150320162259. PMID: 25910510.

[7] Jens H Kuhn and Sina Bavari. *Asymptomatic Ebola virus infections—myth or reality?*. The Lancet.

[8] WHO. *Case definition recommendations for Ebola or Marburg Virus Diseases.* 2014 9 August.

[9] Uyeki TM, Mehta AK, Davey RT Jr., Liddell AM, Wolf T, Vetter P, et al. Clinical Management of Ebola *Virus Disease in the United States and Europe.* The New England journal of medicine. 2016 Feb 18; 374(7):636–46. Pubmed Central PMCID: 4972324. doi: 10.1056/NEJMoa1504874 PMID: 26886522

[10] Caleo G, Theocharaki F, Lokuge K, Weiss HA, Inamdar L, Grandesso F, Danis K, Pedalino B, Kobinger G, Sprecher A, Greig J, Di Tanna GL. *Clinical and epidemiological performance*

of WHO Ebola case definitions: a systematic review and meta-analysis. Lancet Infect Dis. 2020 Nov;20(11):1324-1338. doi: 10.1016/S1473-3099(20)30193-6. Epub 2020 Jun 25. PMID: 32593318.

[11] Andres Colubri, Mary-Anne Hartley, Matthew Siakor, Vanessa Wolfman, August Felix, Tom Sesay, Jeffrey G Shaffer, Robert F Garry, Donald S Grant, Adam C Levine, et al. *Machine-learning prognostic models from the 2014–16 ebola outbreak: data-harmonization challenges, validation strategies, and mhealth applications.* EClinicalMedicine, 11:54–64, 2019 .

[12] Whitmer SLM, Ladner JT, Wiley MR, Patel K, Dudas G, Rambaut A, Sahr F, Prieto K, Shepard SS, Carmody E, Knust B, Naidoo D, Deen G, Formenty P, Nichol ST, Palacios G, Ströher U; *Ebola Virus Persistence Study Group. Active Ebola Virus Replication and Heterogeneous Evolutionary Rates in EVD Survivors.* Cell Rep. 2018 Jan 30;22(5):1159-1168. doi: 10.1016/j.celrep.2018.01.008. PMID: 29386105; PMCID: PMC5809616.

[13] Kerber R, Krumkamp R, Diallo B, et al. *Analysis of Diagnostic Findings From the European Mobile Laboratory in Guéckédou, Guinea, March 2014 Through March 2015.* J Infect Dis. 2016;214(suppl 3):S250-S257. doi:10.1093/infdis/jiw269

[14] Hastie, Trevor, Trevor Hastie, Robert Tibshirani, and J H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer, 2001.

[15] Molnar, Christoph. *Interpretable machine learning. A Guide for Making Black Box Models Explainable* 2019.