Ebola Virus Disease Diagnosis for West African Ebola Virus epidemic

Group eboLAM : Andres Chica Linares, Luis Da Silva, Matthias Zeller

CS-433 - Machine Learning

December 17, 2020

***Abstract*—In collaboration with the IDDO (Infectious Disease Data Observatory, via the University of Oxford) who have expertly curated the biggest Ebola dataset in the world during the 2013-2016 outbreak, we build binary classifiers in order to diagnosis the Ebola Virus Disease (EVD). The automatic quarantine of suspect patients among those infected has resulted in accidental deaths due to the 50% lethality rate of the Ebola virus. Before coming with results, the team had to focus on data cleaning and feature engineering. Unsupervised methods, as Kmeans clustering, PCA and tSNE, have also been performed but without results. Few models could outperform the naive classifiers, however an increased focused on signs and symptoms led us to retrieve most of the symptoms known to be associated with EVD, with an indication of their relative importance. For instance, conjunctivitis and diarrhoea are strongly correlated with EVD. To a lower extent, joint pain, vomitting, asthma and jaundice are also correlated with EVD.**

I. Introduction

Ebola virus has been discovered by the Congolese doctor Ngoy Mushola in September 1976 [3]. This disease is, however, mainly known since the 2013-2016 outbreak in extreme western Africa (especially Sierra Leone, Guinea, Liberia) with a few cases in Europe and USA, leading to 28,646 cases and 11,323 deaths [4]. A less known but higher case-fatality rate outbreak also occurred in Republic Democratic of Congo and Ouganda in 2018 until mid 2020. The Infectious Disease Data Observatory (IDDO) has expertly curated the worldwide biggest Ebola dataset during the 2013-2016 outbreak. During outbreaks, the major problem is that, following arrival of a patient with symptoms, he or she is quarantined with other patients whose diagnosis is also uncertain. As confirmed by OMS [5], Ebola virus disease is hard to distinguish clinically from other infectious diseases such as malaria, typhoid fever and meningitis, only through symptoms. Moreover, only automated or semi-automated nucleic acid testing (NAT) or rapid antigen detection tests is recommended by the OMS to confirm that Ebola infection is the cause of the symptoms [5]. This leads most of the time to global infection in the quarantine zone among all patients. A quarantine could be decided based from the calculated probability, or even use it to sort the patients into different quarantines areas. In general, other accidental infections could be prevented and thus, with the estimated average Ebola case-fatality rate of 50%, deaths could be avoided.

II. Studies

We first focused on a set of curated datasets, consisting of 14 files grouping data from 7 different IDDO Studies (EQJJGF, EJPDEJ, ERFCVU, EOPNOJ, EORKWS, ESYADD, EUZJTB) into different themes. This data was used only to explore our study domain and to grasp the structure of data before focusing into specific studies. Indeed, although the data were supposed to be curated, the data further required extensive cleaning, inconsistencies and missing values were numerous, features and targets were scattered in different files and we would have lost the majority of data by joining those. We thus chose to focus on raw data of individual studies, considering each study as a self-contained dataset and allowed us to have more control on data cleaning.

A. Study 1 : EIXUZQ

The second study ID is EIXUZQ, composed of the patients informations, patients recent contacts, information on some dates (admission death, release, appearance of first symptoms), symptoms, lab tests results and drug administration. Data cleaning, management and exploration, also as a few Machine Learning training models were asked to be performed on this dataset.

B. Study 2 : EGOYQN

The third study ID is EGOYQN, composed of two 2 tabular data being longitudinal clinical observations (784 patients) and triage data (2500 patients). We focused on the triage data, as it contained more patients and the longitudinal format required more data cleaning and processing, partly because some patients had a single observation, whereas others had more than 50 observations. It also contained more features, including basic patient information (sex, age), many indicators of locations, symptoms, contact with other people and dates. The same task was asked for this study as the precedent one, but with an increased focus on the data exploration and cleaning and with additional feature selection.

III. Methods

We used the python Scikit-Learn package to build all our supervised and unsupervised machine learning models, together with standard data science packages for data manipulation and plotting: Numpy, Pandas, Matplotlib, Seaborn. In some cases, we additionally used XGBoost Python package to run our most flexible models and MLXtend package to perform feature selection. We aimed to build binary classifier for the patients epidemiological status, i.e. whether a patient has Ebola Virus Disease or not. Since our goal is primarily inference, we focused on interpretable models that allow to deduce feature importance. We had to focus a lot on data cleaning, but we only mention its key aspect. Moreover, in our case, high recall is preferred over high precision, in order to minimize the false negative rate and prevent people that would not be quarantined to infect other people and spread the virus. Precision is of course also important as we don't want to quarantine healthy people that may become sick because they were in contact with infected people, and may be socially rejected when they come back or die accidentally.

A. Study 1 : EIXUZQ

1) Data Cleaning: The first step was to check any duplicates in the patient ID in the data set, which has to be unique. Then some features were dropped because they were useless in our case or contain too much missing values, more details are present in notebook **??**. All string values have been transformed into lower case format for more homogeneity. The sex feature was transformed into

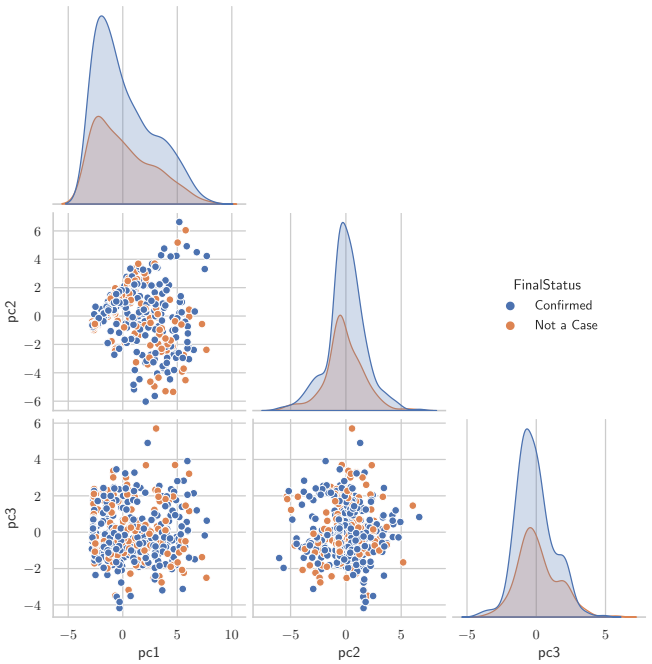

Fig. 1: Pairwise plot of the first three principal components of EIXUZQ study data. All 40 features are included and the three first principal components are shown, colored with Ebola target.

1 for male and 0 for woman. A "referral time" feature, which is the interval between the first appearance of symptoms and the moment of observation, was also created, by substracting (DateIllnessStarted) to (DateofCaseReport). It seems that three tests of Ebola have been made. The values of the second and third report have been changed to NaN values if the first test was not done. Since a lot of studies have been gathered, different terminology have been used, maybe in string values features. Modifications in order to homogenize the data have been made. After cleaning, the features 'Sex', 'Age', 'Referraltime', 'Malaria_positive', 'Malaria_negative', 'FinalStatus', 'TypeOfExit' have been chosen for the next step.

2) Data exploration: We ran some unsupervised models before proceeding to supervised ones. Specifically, we performed Principal Component Analysis (PCA) and KMeans clustering. We included the patient epidemiological status (i.e. Ebola target) in the latter case, as a mean to determine how the data partitioned with respect to the number of clusters and with respect to the target, whereas it was not included in the former case. The first, second and third principal components are depicted in Fig 1, which explain 16.7%, 7.03% and 4.81% of the total variance, respectively. There was no elbow in the plot of ratio of variance explained vs principal components (not shown here) that would allow us to select the optimal number of principal components. In the plot, there is no evident pattern that allow us to conclude that our data can be linearly separable with great performance. The distribution of targets among clusters is shown in Fig. 2. KMeans clustering is not able to partition patients with different targets in different clusters. However, interestingly, the distribution of targets is similar in all clusters, no matter the value of K. Note that for $K > 4$, one cluster is always empty.

B. Study 2 : EGOYQN Triage Data

We obtained a first dataset by keeping only 12 features and obtain as few missing values as possible, before imputing NaNs by most frequent values. This consisted of 10 symptoms, as well as age and

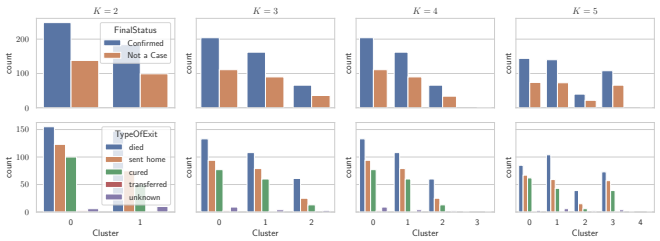

Fig. 2: KMeans clustering of EIXUZQ study data including all features as well as both targets (Ebola and outcome). The distributions of targets is shown for a different number of clusters K.

sex. The second dataset was given to us, containing 105 features and treating missing values in a different way. One-hot encoding was used to encode missing values (e.g. for sex feature, three binary features were created: sex male, sex female and sex unknown), and took advantage of some patterns in the missing values.

Feature selection relied on several methods: stepwise feature selection algorithms (forward and backward selection), selection by absolute value of coefficients for a ridge classifier model, and selection by p-values. Eventually, we could combine those methods by taking the intersection or the union of features selected by each method.

In order to increase models performance, we extended the feature space by computing all combinations of pairwise-multiplied features. This procedure allows to capture pairwise interactions between features, and thus increase flexibility of our linear model. Note that the number of such generated features grows as $\Theta(n^2)$, thus upstream feature selection is critical to avoid fitting models on a ridiculously high number of features (with respect to the number of samples).

In order to decrease models variance, for each binary feature, we compute the number of patient falling in each category, and drop the feature if either category has less counts that a chosen binary feature filtration threshold T. T may be interpreted as a regularization hyperparameter. This is especially useful — even necessary — to combine with the aforementioned feature expansion, as many features of extended feature space will contain very few patients per category.

1) Data exploration: We also ran unsupervised learning methods for this dataset. The plot of explained variance ratio vs principal components (not shown) suggested to keep the first two principal components (over 83), accumulating 21.0% of the total variance. However, plotting the samples projected on the first two principal components and coloring with the target did not exhibit any interesting pattern, tSNE visualization neither. For K-means, the plot of the sum of square distances to the closest cluster center vs number of clusters did not show any elbow.

IV. Results

A. Study 1 : EIXUZQ

The dataset was first splitted into training and test set. 10-fold cross validation was used to tune some model hyperparameters, under several scoring metrics (accuracy, precision, recall).

As can be seen in Table I, none of the models we trained significantly outperforms the naive classifier. Indeed, the best model (k-NN) has 10% less recall but only 6% more precision, compared to the naive model. Feature importance of the ridge regression model is shown in Fig. 3.

When fitting a logistic regression on 24 features, and filtering out those associated with a p-value greater than 5%, we end up with

TABLE I: Results for EIXUZQ study.

Training model	Accuracy	Recall	Precision
Naive Classifier	65.05	100.0	65.05
Stochastic Gradient Descent	57.31	56.94	75.79
Ridge Regression	61.94	51.76	81.48
Support Vector Machine	68.02	90.69	68.85
K-Neighbor	70.99	90.83	71.64
Nearest Centroid	64.29	57.08	81.19
Gaussian Process	66.54	87.36	68.98
Decision Trees	68.63	83.75	71.9
Random Forest	67.31	86.25	69.86
XGBoost Classifier	69.4	89.41	70.37
Random Classifier	53.73	50.59	68.25

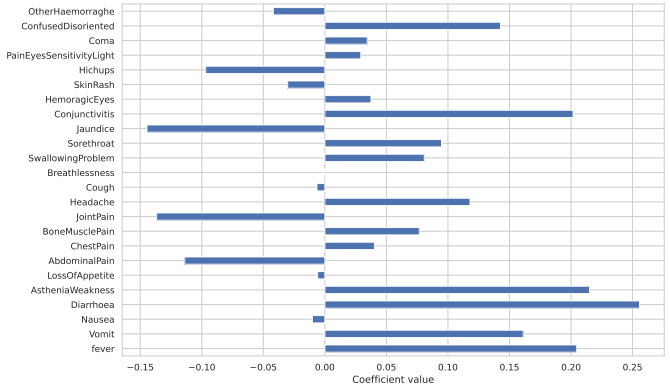

Fig. 3: Feature importance of ridge classifier model on 24 symptoms for EIXUZQ study data.

6 features, being jaundice, joint and abdominal pain, swallowing problems, diarrhoea and conjunctivitis. This model suggests that the most strongly associated symptoms are Jaundice, with an odds ratio (OR) of 0.779, diarrhoea (OR 1.23) and conjunctivitis (OR 1.27), which is consistent with feature importance obtained with coefficients of the ridge classifier. Although jaundice may be a sign of both ebola virus disease [2] and malaria [1], it rather suggests an absence of ebola disease in our dataset. This is also confirmed by the strongly negative coefficient of jaundice for the SGD classifier.

B. Study 2 : EGOYQN, Triage data

Starting from the approximately-balanced dataset of 12 features, we ran logistic regression and obtained p-values to filter statistically non-significant features. Removing non-significant features actually reduced the overall performance of the model, as shown in part (a) of Table II. Features with biggest effect on prediction are diarreha, vomit and hemorragy symptoms, having odds ratio (OR) of 2.31, 1.94 and 1.82, respectively. Including interaction terms increases recall, at the expense of precision and accuracy. Model with interactions assigns an unexpectedly strong effect to hiccups (OR 5.76) towards positive diagnosis, which is largely counterbalanced if patient has joint pain or anorexia (interaction hiccups-joint pain has OR 0.121, hiccups-anorexia has OR 0.225). Having both headache and asthma also strongly suggests positive diagnosis (OR 4.15).

Starting from the other dataset of 105 features, we identified 18 binary features of the training set having less than the threshold of $T = 10$ patients for either category, and removed those from training and test set. Note that this dataset has strongly unbalanced labels, as indicated by the naive classifier of Table II. We ran stepwise forward and backward selection, as depicted in Fig 4, which both reach a plateau for more than about twenty features, with a significant variance when using more than 30 features. We decided to extract

TABLE II: Summary of test set results for triage data of EGOYQN study

Accuracy	Recall	Precision	Note
(a)			
58.7	100.0	58.7	Naive classifier
64.9	71.0	69.8	Logistic regression
64.5	65.2	71.7	Logistic regression, p-value filter
65.1	78.5	67.4	Regularized logistic regression, interaction terms
(b)			
66.4	100.0	66.4	Naive classifier
71.3	83.3	75.4	Ridge classifier
71.3	83.3	75.4	Ridge classifier, interaction terms
73.6	92.6	74.1	Logistic regression, p-value filter
76.2	86.6	80.0	Bagging with decision trees, interaction terms

P-value filtration refers to remove features with p-value greater than 5%. Interaction terms refers to pairwise-multiplied features. **(a)** Dataset of 12 features (few symptoms, age, sex). **(b)** Dataset of 105 features.

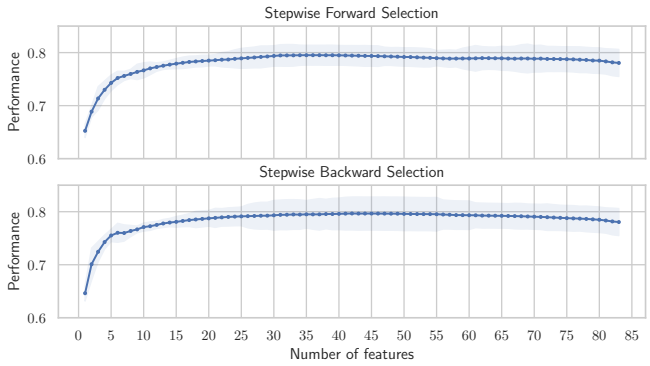

Fig. 4: Stepwise feature selection for EGOYQN study dataset of 105 features using ridge regression model. Shaded area represents 95% confidence interval computed on five folds.

two sets of the best thirty features of each method. We also obtained the thirty best features of a baseline ridge regression model, based on feature coefficients. Some results are summarized in part (b) of Table II.

We ran sets of ridge regression models with different datasets, as shown in figure 5. All four models are equivalent in F1 score terms under appropriate regularization, the difference lying in the tradeoff between precision and recall. With those results in hands, we performed similar cross-validated analyses with ridge regression and support vector classifier models, on both the union and the intersection of the three sets of thirty features, and by modifying the filtration threshold T of interaction features. We obtained similar results.

Alternatively, fitting an L1-regularized logistic regression model on all features and filtering out features associated with p-values $<$ 5% yields good performance (see Table II) for only ten features. Among remaining significant features, conjunctivitis has a very strong positive effect (OR 13.2) — a known sign of EVD [6] — while retro-orbital pain has a strong negative effect on ebola diagnosis (OR 0.122). Note that the features selected by all our methods are asthma, conjunctivitis, diarrhoea, hemorrhage, hiccups, retro-orbital pain, vomiting, and a feature very specific to this pandemic being the day of year (between 1 and 365) for the arrival of patients at the Guéckédou hospital.

The last step was relax the interpretability constraint and perform similar analyses on a bagging classifier, with decision trees and support vector classifiers as base estimators. The best test-set result we achieved was with decision trees, including interaction terms.

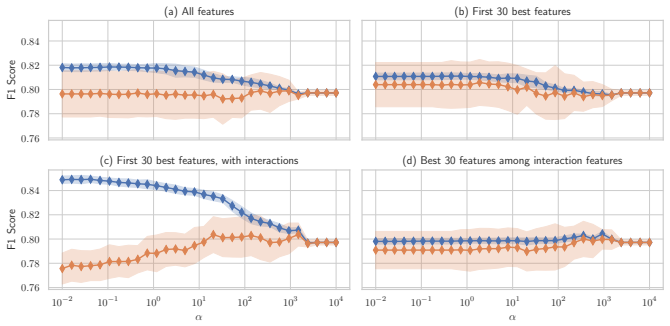

Fig. 5: Regularization hyperparameter cross-validation of ridge regression model for EGOYQN study data. A higher α value reflects stronger regularization. F1 score is used as the cross-validation metric. **(a)** All 83 features as a comparator model. **(b)** Best 30 features of (a), selected by absolute value of coefficients. **(c)** Features of (b) are multiplied pairwise and non-informative resulting features are dropped, ending with 301 features (binary feature filtration threshold $T = 10$). **(d)** Best 30 features of (c), selected by absolute value of coefficients.

V. Discussion

Although some of our models yield some contradictions, as a negative association between hiccups and the disease in the ridge classifier of study 1, whereas we found the opposite in study 2, some similar conclusions can be drawn. For instance, conjunctivitis and diarrhoea are strongly correlated with ebola in the majority of our models. To a lower extent, joint pain, vomitting and asthma are also (positively or negatively) correlated with EVD.

It is naturally straightforward to correctly diagnose patients infected by EVD as positive (i.e. having 100% recall), as a naive classifier would do. Most of the attempted methods rather consisted to adjust the trade-off between recall and precision, as few models can be claimed to outperform the naive classifier, although the difference is small. Indeed, the constraint of interpretability naturally limited the performance we could obtain, given our dataset. This also shows how critical the data cleaning procedure is, as there is no model that is independent of the quality of input data. We speculate that more extensive data cleaning and a deeper understanding of the data collection performed in the field (e.g. through close collaboration with data collectors) may have allowed us to further exploit the available data. Moreover, some future work may take advantage of the numerous datasets collected by different studies, to increase the total number of patients, build more flexible models and achieve better performance.

VI. Conclusion

Learning methods were not totally able to predict with great accuracy, although focusing on signs and symptoms did led us to retrieve most of the symptoms known to be associated with EVD, with an indication of their relative importance.

Nonetheless, this project has allowed us to discover the importance of data cleaning and feature management beforehand applying training models, especially in uncleaned, longitudinal data with a lot of features, thus confirming once again the the Pareto Principle of data analytics and machine Learning : 80% of the time is devoted to data preparation and the rest of the process is about real analysis and models fine-tuning.

Evaluating training methods was also complicated by the fact that there is no absolute scoring methods that allows us to clearly rank the methods among them, as it depends on what goal we want to achieve.

The size of the datasets was also an obstacle when applying the models. For the first study we had 669 samples and for the second 1700. We hypothesise that with a bigger data we could possible achieve more significant results.

Eventually, the concluded observations made upon this EVD modelling study could be reused later in case of a new Ebola outbreak. For example, a reduction in the number of missing values, which in the field would result in more complete completion of health forms, despite the professionals in the field being immensely solicited during the outbreak.

VII. Acknowledgments

We thank Mohammed Ridha Chahed for his help and his availability during the project, the team's work may have been inspired by his notebooks, the IDDO Ebola Data Repository as well as all the organisations for their datas collection and sharing : Alliance for International Medical Action (ALIMA), International Medical Corps (IMC), Institute of Tropical Medicine Antwerp (ITM), Médecins Sans Frontières (MSF), Oxford University and Save the Children (SCI).

References

[1] Anil C. Anand and Pankaj Puri. "Jaundice in malaria". In: *Journal of Gastroenterology and Hepatology* 20.9 (Sept. 2005), pp. 1322–1332. ISSN: 0815-9319. DOI: 10.1111/j.1440-1746.2005.03884.x.

[2] Nicholas J Beeching, Manuel Fenech, and Catherine F Houlihan. "Ebola virus disease". In: *The BMJ* 349 (Dec. 10, 2014). ISSN: 0959-8138. DOI: 10.1136/bmj.g7348. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4707720/ (visited on 12/16/2020).

[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". In: *The Journal of Infectious Diseases* 214 (Suppl 3 Oct. 15, 2016), S93–S101. ISSN: 0022-1899. DOI: 10.1093/infdis/jiw207. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5050466/ (visited on 12/16/2020).

[4] Cordelia E. M. Coltart et al. "The Ebola outbreak, 2013–2016: old lessons for new epidemics". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1721 (May 26, 2017). ISSN: 0962-8436. DOI: 10.1098/rstb.2016.0297. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5394636/ (visited on 12/17/2020).

[5] "Maladie à virus Ebola". fr. In: (). URL: https://www.who.int/fr/news-room/fact-sheets/detail/ebola-virus-disease (visited on 12/07/2020).

[6] Majid Moshirfar, Carlton R Fenzl, and Zhan Li. "What we know about ocular manifestations of Ebola". In: *Clinical Ophthalmology (Auckland, N.Z.)* 8 (Nov. 21, 2014), pp. 2355–2357. ISSN: 1177-5467. DOI: 10.2147/OPTH.S73583. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4247133/ (visited on 12/15/2020).