Biostatistics

Prediction of Bleeding Episode Among Vitamin K Antagonist Treatment Patients

Eduardo Mendes (76091), Bruna Mason (78686), Liana Mehrabyan (91199),
Léo Gire (91238), Bastien Gros (91374)
Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal

***Abstract*—Vitamin K Antagonist (VKA) treatment, performed through oral anticoagulant drug perscription, crucially relies on the envisioning of possible unwanted episodes that the patients might have. A dataset of Portuguese VKA patients has been used to predict the bleeding events that may occur. A preliminary analysis has been conducted on the dataset which was followed by applying supervised learning algorithms, most importantly, logistic regression.**

Keywords: Vitamin K Antagonists, bleeding, healthcare, Supervised Learning, Logistic Regression, Support Vector Machines, K-nearest neighbours, Random Forests, Relief, SMOTE.

I. Background

Vitamin K antagonists (VKA) are a group of substances that reduce the action of vitamin K. They are used as anticoagulant medications in the prevention of thrombosis as well as pest control, as rodenticides. The VKA treatment is very often accompanied by bleeding events and it is very useful to be able to predict and prevent those events. A dataset of 324 Portuguese patients' medical data was provided in order to develop models that will predict the event of bleeding among the patients that receive VKA treatment. The International Normalized Ratio (INR) is a laboratory measurement of how long it takes blood to form a clot and it is the best measure to evaluate the efficacy of the drug. The target range for INR values is between 2.0 and 3.0. Low values of INR are associated with a higher risk of developing a clot, while higher values of INR indicate a higher risk of bleeding. For each patient, regular INR checks are scheduled where, if the INR is out of range, a different VKA dosage will be administered. This is due to the fact that VKA has an important inter-individual and intra-individual variability. In order to estimate these values for the rest of the days, it is used the linear interpolation method of Rosendaal, which assumes a linear variation of INR values between visits. The time in therapeutic range (TTR) is used as a measure of quality of the anticoagulation control. Values of TTR higher than 70% are associated with a good quality of anticoagulation control, while values of TTR lower than 60% are related to adverse events.

A total of 44 features were provided by the dataset, most important of which include: total days of VKA treatment; TTR variables concerning the time in therapeutic range; variables concerning the values of INR; patients' age and gender; several clinical outcomes such as strokes, heart failures and myocardial infarction, as well as CHADSVASC: a risk score for thromboembolic events that evaluates the previous clinical characteristics – age, heart failure, hypertension, diabetes, stroke, vascular disease and female gender. Not all of the variables were included in the analysis as there were variables which were not informative for this particular project having bleeding prediction as its main objective. The overall procedure and reasons for eliminating several variables will be described in further sections of this paper. The analysis was performed using R software.

II. Data Pre-Processing

One of the main challenges of the project is having highly imbalanced data (only 33 bleeding episodes out of 324 observations). The other challenge was to identify variables that were not informative for the problem from the medical point of view. To tackle these problems in the best way possible, the original dataset had to suffer some modifications. First off, since the goal is to predict a bleeding episode it wouldn't make sense to include the other episodes (stroke, death, MI) that occurred after the bleeding. Having verified that only for 4 observations was there a non-bleeding event previous to a bleeding one, it was concluded that it would also be insignificant to study for the events that happened beforehand. Accordingly, it was decided to remove the variables related to the non-bleeding episodes. Since the dates would only be important to the problem at hand to discriminate if the non-bleeding events happened before or after the bleeding event, these variables were also removed from the dataset. The *FUP* and the *Morte CV* variables were also removed since the last known follow-up date and the cause of death has no importance to the problem.

The only values that were necessary to edit from the original dataset were the *NA* values for the *No. Tests in Range* variable. These *NA* values were switched for zeros after making sure with the person responsible for this dataset that a *NA* value corresponded to a no. of tests in range equal to zero.

At this point, it was created the first dataset which includes all the original variables except the ones we mentioned previously. This dataset was named **data**. It was also created a new dataset, called **datasd**, where the numerical variables were standardized according with the standard normal distribution.

A principal component analysis (PCA) was conducted to deal with the highly correlated variables which might lead to

misleading results for some of the methods that were applied. This analysis is only valid for the quantitative variables and it returns a new set of variables called principal components. PCA transforms the existing variables by a simple linear transformation into principal components that are uncorrelated with each other. This analysis was performed for the standardized dataset. Only the first four principal components were selected for posterior analysis since, together, they explained 81.66% of the cumulative proportion of variance. Afterwards, a new dataset was created, called **datapca**, where this 4 components were merged with the remaining categorical variables.

For validation purposes, these datasets were divided into two sets, preserving the original class percentage of the target variable in both of them: 70% of the observations were assigned to training the models and 30% of the observations were assigned for testing (**data.test**). As a result the following datasets were defined: **data.train**, for the original dataset, **datasd.train**, for the standardized dataset, and **datapca.train** for the pca dataset.

On an initial intuitive analysis, for the variables related with the INR being greater or less than a certain value, the number of days this happened for a certain patient should be more relevant than its respective percentage. This is because a high percentage could be obtained for an individual whose participation in this study was very short, which should not be associated with an equally high risk of occurrence of an adverse event. It was observed that for the number of tests in range or with a value higher or lower than a certain threshold, only the respective percentages were provided for each patient. Being so, it was decided to try to check if, by adding new variables for each case where the number of tests an individual has done in a certain range was included, more relevant information could be added. However, when including these variables in the different methods, no significant results were obtained. Thus, it was decided to omit these results from the study.

Since the original dataset is unbalanced for the response variable, some measures had to be taken to deal with this problem. More precisely, it was performed undersampling, oversampling and both undersampling and oversampling to the data, to assure that the proportion of observations for the both cases (having a bleeding or not) would be approximately the same. The oversampling approach replicates the observations coming from the minority classes until all classes are even, while the undersampling approach removes observations from the majority classes. Both of these approaches can be applied by removing observations from the majority class and replicating observations from the minority class simultaneously. An additional, Synthetic Minority Oversampling Technique (SMOTE) was applied to the data as well. The algorithm takes the differences between the observation under consideration and its k nearest neighbors. It then multiplies these differences by random numbers between 0 and 1, adds it to the feature vector under consideration thus generating k synthetic observations cite1. These methods (over, under, both and SMOTE) were only applied to the corresponding training set of each previously mentioned dataset, resulting in the following datasets: **data.trainOver**, **data.trainUnder**, **data.trainBoth**, **data.trainSMOTE** (for the original dataset); **datasd.trainOver**, **datasd.trainUnder**, **datasd.trainBoth**, **datasd.trainSMOTE** (for the standardized dataset); **datapca.trainOver**, **datapca.trainUnder**, **datapca.trainBoth**, **datapca.trainSMOTE** (for the pca dataset).

III. Preliminary Analysis

The initial dataset, after the data pre-processing stage was carried out, contains 35 features of INR records, clinical characteristics and outcomes of 324 Portuguese patients that received VKA treatment. In order to study the population's characteristics a descriptive analysis was conducted on the variables considered the most relevant by computing ranges, means and standard deviations. This sample of patients consists of 60% males and 40% females, whose ages vary between 24 and 92 years old (as shown in Figure 1), 39% of which have an age greater or equal to 75, which carries a high risk of occurring a cardiovascular-related event. Most of the clinical conditions have imbalanced classes, except for the *Heart Failure* feature. *Diabetes* and *Stroke* conditions have really imbalanced classes, 20% to 80%, while *Hypertension* and *Vascular* conditions have a proportion of 67% to 33%. The range of the total number of days spent in treatment varies between 63 days and 1001 days, with a median equal to 405 days, meaning that half of the patients spent less than 405 days in treatment while the other half spent between 405 and 1001 days. As for the number of tests each patient undertook, the values vary between 6 and 47, with a median equal to 17 tests performed. The distribution of the *Chadsvasc* scores among the patients is represented in the Figure 2. It can be observed that most of the individuals have a score between 2 and 5, which indicates a moderate risk of suffering a thromboembolic event.

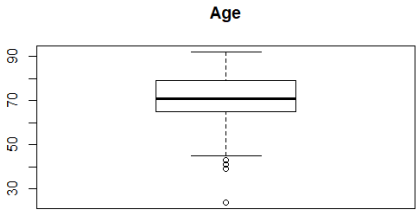


Fig. 1: Boxplot of *Age* variable

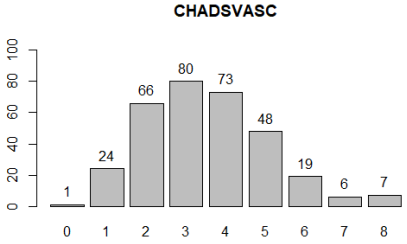
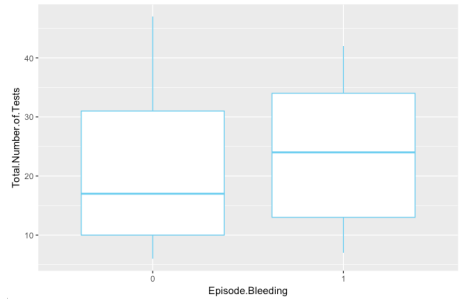
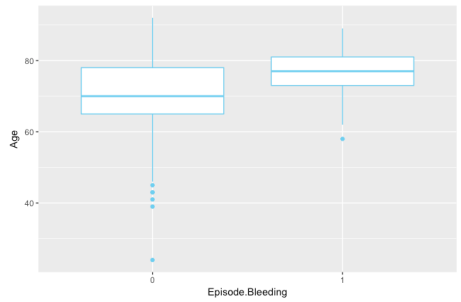


Fig. 2: Barplot of *Chadsvasc* variable

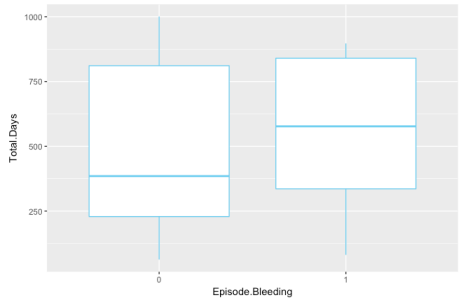
In order to determine which variables are associated with the occurrence of the bleeding episodes, the dataset was split into two groups: one with the patients who suffered a bleeding episode and the other with the remaining patients who did not experience such an event. For the quantitative variables a Wilcoxon Signed-Rank Sum Test was performed, through the *wilcox.test* function, to evaluate if the means were the same or not for both groups. For this hypothesis test, the null hypothesis, H_0, states that the means for the variable tested in the two groups is the same. By rejecting the null hypothesis, there is a strong evidence that the means are different between each group, which indicates that the variable in study is significantly important to discriminate the event. Therefore, at a 5% significance level, H_0 was rejected for the following variables: *Total Days*, *No. Days w/ INR<2*, *No. Days w/ INR>3*, *No. Days w/ INR<1.8*, *Total Number of Tests* and *Age*. Boxplots of the variables also confirm the results obtained by the test, in Figure 3 are represented some of these plots.



(a) *Total Number of Tests*



(b) *Age*



(c) *Total Days*

Fig. 3: Boxplots of some quantitative variables, comparing the non-bleeding subset with the bleeding subset

For the categorical variables, a Pearson's Chi-Squared Test was performed, through the *chisq.test* function, to check what variables are independent of the response variable. The H_0 of this test states that both variables are independent. Thus, rejecting H_0 means that the variables are dependent, i.e., the variable at test should have an influence on the outcome variable. Hence, at a 5% significance level, H_0 was rejected for the following variables: *Age≥75*, *Vascular* and *Chadsvasc*.

In order to verify and expand the results of the test, Relief feature selection algorithm was applied by *attrEval* function. The algorithm was originally created for binary classification problems with discrete or numerical features which exactly fits current data under inspection. It calculates a feature score for each feature by calculating feature value differences between nearest neighbors [2]. The results of the algorithm confirmed the ones from the former hypothesis tests and marked a new variable, *Days Within Range*, to be significant.

Considering that the correlations of the features may be medically important, the correlation matrix illustration is presented in Figure 4. Through its observation it is possible to verify that a considerable amount of the variables present in the dataset are highly correlated with each other. For instance, the *Days Within Range* variable presents a high correlation with the *Total Days*, *Total Number of Tests* and *Number of Tests in Range* covariates, while the *No. Days w/ INR<1.8* is highly correlated with *TTR*, *% Days w/ INR<2*, *% Days w/ INR<1.5*, *% Tests w/ INR<2* and *% Tests w/ INR<1.8*. However, surprisingly enough, *Age* is not highly-correlated with any other features. The large number of correlated covariates is easily explained by the fact that most of these measures were calculated using the information present in some of the other variables. Moreover, some variables are just a representation of the percentage from a quantity that is included in another covariate. Thus, creating models that contain all of these variables should not generate good results. Therefore, reducing the number of variables to build the models and ensuring that there is no high correlation between the remaining variables was essential for this project.

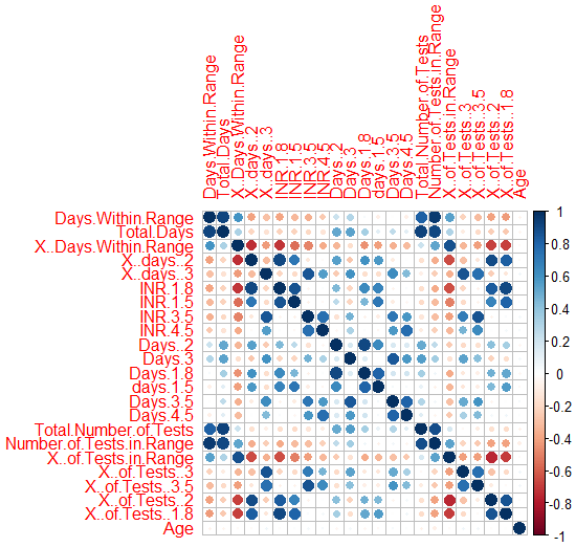


Fig. 4: Correlation of Numerical Variables

The exploratory analysis gave some insights about what variables may be important to explain the outcome variable. This knowledge will be used in the construction of some models.

It was decided to use Youden's index and AUROC as metrics of interest in the selection of the best models, as they present more cumulative results in terms of sensitivity and specificity.

IV. Methods

A. Logistic Regression

In order to describe the relationship between the bleeding episodes and the input variables in the best way possible, a Logistic Regression model was built. The different fitted models were tested for all the datasets previously mentioned.

Before any analysis, it was decided to consider, for this method, the *Chadsvasc* variable as a continuous variable and not as a categorical one because *Chadsvasc* is a score and does not represent a specific characteristic. For example, there is a link between the value of the score you get : to get a score of 4, you had at a time a score of 2 or 3, and then reached 4.

In the first place it was conducted an univariate analysis to identify the important variables, where the ones with a p-value higher than 0.25 were removed. After this step, the number of variables were reduced from 34 to 18: *Hypertension, Stroke, Vascular, Heart Failure, Female, Age≥75, TTR, Total Number of Tests, No. Days W/ INR>3, No. Days w/ INR<2, No. Days w/ INR<1.8, No. Days w/ INR>3.5, % Days w/ INR>3, TTR<60%, TTR>75%, % Days Within Range, Total Days* and *Chadsvasc*. Although the p-value of the *Age* variable was larger than 0.25, this feature was also kept in the model due to clinical considerations. Afterwards, it was created a multiple logistic regression model with all the 19 variables previously selected. Subsequently the variables with a larger difference between the uni- and the multivariate estimates were removed, ending up with 12 variables. In addition, the *No. Days w/ INR<1.8* and *No. Days w/ INR<2* variables were also removed since their p-values were very high.

After a quick analysis of the variables, it was observed that the current model contained 4 variables (*Vascular, Stroke, Heart Failure* and *Female*) used to calculate the *Chadsvasc* score. After a comparison between this current model and the one where these variables were replaced by the *Chadsvasc* variable, it was decided to keep the second one because it simplified the model and did not affect the performance. Finally, it was obtained a model with 6 variables: *Total Number of Tests*, *No. Days w/ INR>3.5*, *No. Days w/ INR>3*, *TTR<60%*, *Total Days* and *Chadsvasc*.

Since the Logistic Regression method doesn't perform well with correlated predictors, one of these two variables, *No. Days w/ INR>3* and *No. Days w/ >3.5*, had to be removed since they are highly correlated. It was decided to keep the *No. Days w/ INR>3.5* variable since the accuracy of the model was higher when this variable was kept instead of the other.

The next step was to evaluate the linearity assumption for the continuous covariates of the logit for the continuous variables. The smoothed scatter plots, represented in the Figure 5, show that variables are all quite linearly associated with the *Episode Bleeding* outcome in the logit scale, so it's not necessary to perform any kind of transformations on the covariates.

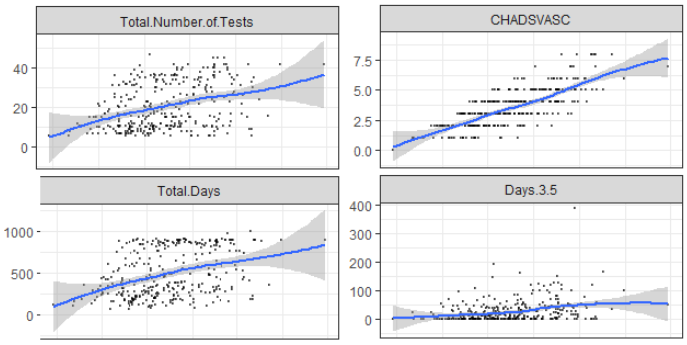


Fig. 5: Smoothed scatter plots (logit against variables)

Furthermore, to calculate the predicted values, it was determined a cut-off, using the *optimalCutoff* function, to maximize the Youden's index.

The last step was to verify the existence of interactions in the remaining variables of this model. After an analysis of those interactions (creation of a list of possible interactions and verification of their significance using the likelihood ratio test), it was found a logic interaction between the variables *Total Number of tests* and *Total Days*. So, this interaction was added to our model and subsequently tested in our **data.test** to check whether there was an improvement or not. Afterwards it was also performed a Wald test, that guaranteed the significance of this interaction. The comparison between the models with and without the interaction term is comprised in the Table I.

	Model without int term			Model with int term		
Accuracy	0.5208			0.8438		
Sensitivity	0.7778			0.7778		
Specificity	0.6437			0.8391		
Youden's index	0.4215			0.6169		
AUROC	0.751			0.8282		
confusion matrix		reference			reference	
	prediction	0	1	prediction	0	1
	0	56	2	0	73	2
	1	31	7	1	14	7

TABLE I: Influence of the interaction term

Since the best results were obtained for the model with the interaction term, the final model is the one with the following variables: *Total Number of Tests*, *No. Days w/ INR>3.5*, *TTR<60%*, *Total Days*, *Chadsvasc* and *Total Number of Tests * Total Days* (our interaction term).

The model that includes the interaction factor provided better results, since the value of the specificity increased while the sensitivity remained the same, meaning that the model was able to reduce the misclassification happening with the majority class.

Another way to build a model is to run an automatic stepwise selection procedure which is only based on statistical significance. But, we have to be careful with this procedure

because it can lead to biologically implausible models. This procedure has to be checked at every step.

It was run in two different ways:

- the forward selection one in which we start with a simple model and we add the most significant term at each iteration until any further additions don't improve the fit of the model
- the backward elimination in which we start with the new model with all the variables kept after the univariate analysis and removed from this model a variable at each iteration until another remove leads to a lost in the fit of our model.

These models were run on the original, oversampled and undersampled data and two of the data gave the best predictive results. In the first model obtained on the **data.train** dataset with a forward procedure, the following variables were included: *% Days with INR>3*, *Age*, *Vascular*, *Stroke* and *Days Within Range*. And in the second model, obtained on the **data.trainUnder** dataset with a backward procedure, the following variables were kept: *Total Number of Tests*, *No. Days w/ INR<1.8*, *No. Days w/ INR<2*, *No. Days w/ INR>3*, *Total Days* and *Chadsvasc*.

The results are summarized in the Table II.

	Model 1			Model 2		
Data	data.train			data.trainUnder		
Accuracy	0.875			0.8958		
Sensitivity	1.0000			0.7778		
Specificity	0.4712			0.8161		
Youden's index	0.4712			0.4828		
AUROC	0.7292			0.7318		
confusion matrix		reference			reference	
	prediction	0	1	prediction	0	1
	0	41	0	0	71	3
	1	46	9	1	16	6

TABLE II: Stepwise Logistic Regression results

It can be seen that the first model is able to predict all the future bleeding events, since it has a sensitivity of 1. However many false positives are present in the results. With our second model, the specificity is way better (82%) though there was a decrease in the sensitivity measure (78%). Although the second one has higher results of Youden's index and AUROC, the first one could be kept by medical considerations in order to create a group of risky patients and not miss any patient that could bleed in the future.

Still, the best model obtained is the one which was generated by the univariate analysis with the interaction term, since it has a Youden's index of 0.6743 and AUROC of 0.8448. This model will be kept as our best model obtained with the logistic regression.

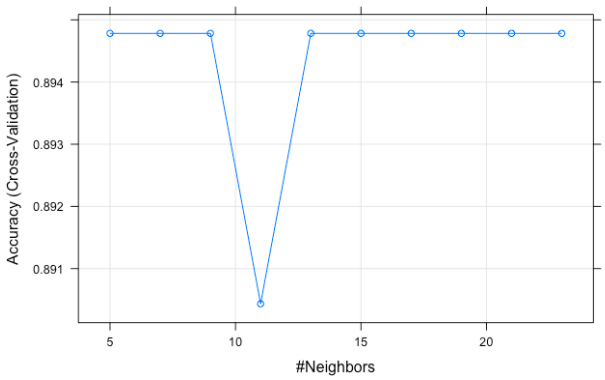
With this model it was calculated the odds ratio (OR) of several variables:

- for the *TTR<60%* variable, the OR was 2,465, meaning that, when a patient was less than 60% of the time within the therapeutic range, his odds of having a bleeding event increase by 147%. This is according to our initial intuition for this feature, since when, for a patient, the coagulation control is not adequate, it should increase the chances of occurring a bleeding;
- for the *Chadsvasc* variable, the OR is 1.5679, so when a patient gains one point on his *Chadsvasc* score his chances of having a bleeding event increase by 57%. This covariate, from a medical point of view, should favour the occurrence of non-bleeding events instead of the bleeding ones. This is due to the fact that it measures the risk of thromboembolic events, which are related to the creation of clots in the blood. Thus, these results go against this medical interpretation.
- as for the *No. Days w/ INR>3.5* variable, the OR was 0.731784 for a period of 30 days, meaning that if the INR of a patient is over 3.5 during one month more, the odds of having a bleeding episode decrease by 27%. This is totally contrary to the clinical perception of this feature. Higher values of INR should always indicate a higher risk of a patient suffering from a bleeding event, since the blood would be more fluid. It makes no sense whatsoever that it should favour the occurrence of non-bleeding events over the bleeding ones.

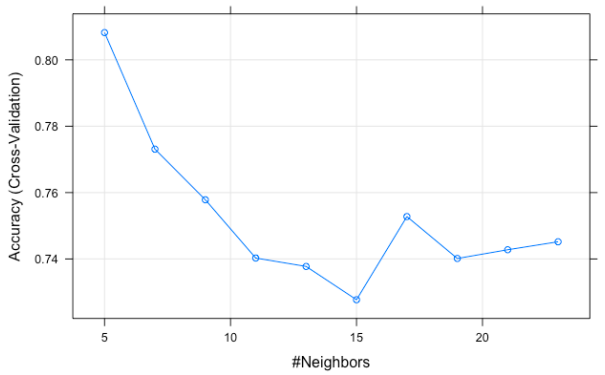
These contradictory results may have been originated by some kind of computational error when applying the logistic regression model or maybe, although an initial medical interpretation and intuition about these feature lead us to certain conclusions, when executing a classification method such as this one on real data, different inferences could be achieved that make sense on a mathematical point view but not on a medical one.

B. KNN

In order to compare the performance of Logistic Regression model other Supervised Learning algorithms were applied to the data. The first algorithm to be presented it the K-nearest neighbors (KNN) classification algorithm. KNN calculates the distances between the case to be classified and all other observations, chooses the k nearest neighbors and assigns the class that dominates among the k neighbors. There are various ways to calculate the distances such as the Euclidean, Minkowski, Manhattan, Chi Square and others, however Euclidean distance was used for this data. Categorical variables of the data were treated as numerical, which is a common practice. The optimal number of k neighbors was chosen by cross-validation parameter tuning and gave different results for different preprocessed datasets that will be discussed later. KNN was first applied on the initial dataset without over or under sampling. After applying exploratory data analysis and feature selection, a total of six variable were used in the final version of the algorithm: *Total Days*, *Total Number of Tests*, *Age*, *Age≥75*, *Vascular* and *Chadsvasc*. The algorithm was applied on two datasets: the original one, **data.train**, and a dataset obtained by SMOTE oversampling, **data.trainSMOTE**, in order to handle the imbalance issue. As shown in the Figure 6, the parameter tuning gave the same optimal number of neighbors to be $k = 5$.



(a) Original Data



(b) SMOTE Data

Fig. 6: Accuracy depending on Number of Neighbors

The results are presented in the Table III.

	data.train			data.trainSMOTE		
Accuracy	0.9375			0.9167		
Sensitivity	0.33333			0.7777		
Specificity	1.00000			0.9310		
Youden's index	0.3333			0.7087		
AUROC	0.9859			0.9719		
confusion matrix		reference			reference	
	prediction	0	1	prediction	0	1
	0	87	6	0	85	0
	1	0	3	1	2	9

TABLE III: KNN Results

Choosing a winner between these two results depends on the criterion of interest from medical point of view. KNN on SMOTE treated dataset gave better results Youden's index and sensitivity but a bit lower results in terms of accuracy and AUROC. Training and testing validation errors were also calculated and proved the absence of over fitting. There are many opinions whether oversampling is a valid tool and does not create bias in the data and our opinion about oversampling is positive as SMOTE algorithm does not create any duplicate observations and oversampling is done more objectively. Note, that all the variables were later converted to integers after SMOTE treatment in order to have the same structure of the original data. It is worth to note that apart from these two experiments, KNN algorithm was also applied to PCA-processed data and under sampled data, however, for the sake of conciseness, it was decided to discuss the two best results only. One of the advantages of KNN is its easily interpretable idea lying behind the algorithm. KNN has a high interpretable value in medical problems as patients with similar conditions usually confront similar events. Therefore, having known what events confronted patients having medical records the most similar to the patient under interest can predict the episodes that will happen with considerable preciseness. A future improvement of this method would be using chi.square distance instead of Euclidean distance. As suggested in [2], usage of KNN algorithm on mixed medical data of categorical and numerical variables performs the best when using chi.square distance.

C. SVM

The next Supervised Learning method applied for this problem was the Support Vector Machine (SVM). For this method, each observation will be attributed a set of coordinates in a n-dimensional space (n is the number of variables at study) according to the values of each of its variables. The goal of this algorithm is to find the hyperplane or set of hyperplanes that best separate the points belonging to each class. This is achieved by maximizing the distance between the hyperplane and the nearest point belonging to each class, thus obtaining the maximum margin possible. New observations are plotted in that same n-dimensional space and then they are classified according to which side of the hyperplane they belong.

The most important parameter when employing this method is the C value, which determines the size of the margin of the constructed hyperplane. For large values of C, the algorithm will build smaller-margin hyperplanes separating the classes, while for low values of C, higher-margin hyperplanes will be produced. We are looking for a compromise between over fitting the model with a small margin and eventually increasing the misclassification error with a larger margin. There is no rule of thumb on how to choose the best value of C for a dataset, so for each simulation it is essential to perform cross-validation within the training set in order to find the optimal C, i.e. the one that minimizes the error rate. In our case, this was executed for each simulation with the *train* function in R, setting a grid with values of C between 0 and 2.5, which returns a model with the optimal C that later is used for prediction.

It is also important to observe that the building of the hyperplane for the SVM algorithm can be accomplished by using either linear or non-linear methods. Both procedures were tested, but reasonable results were only obtained for the linear case. So, the radial approach for the SVM algorithm was omitted for the following results.

For this method the best results for the Youden's Index and AUROC were obtained for the model produced by the **datapca.trainBoth** dataset considering the 4 principal components and the *Chadsvasc* variable. The second best results were generated by the **datasd.trainSMOTE** for the variables *Total Days*, *No. Days w/ INR<2*, *No. Days w/ INR>3*, *Age*,

Vascular and *Chadsvasc*. These results can be observed in the Table IV.

	datapca.trainBoth w/ 4 princ. comp. + Chadsvasc	datasd.trainSMOTE w/ chosen variables
Accuracy	0.7604	0.7083
Sensitivity	0.77778	0.77778
Specificity	0.75862	0.70115
Youden's index	0.5364	0.47893
AUROC	0.7682	0.7395

confusion matrix

prediction \ reference	0	1
0	66	2
1	21	7

prediction \ reference	0	1
0	61	2
1	26	7

TABLE IV: SVM Results

The values obtained for the accuracy and Youden's index with these models were not as high as it would have been desired. Even though the sensitivity reached relative high values taking into consideration the low amount of positive events, the specificity would have needed to be slightly higher in both cases in order to decrease the misclassification rate, which was quite high. These somewhat disappointing results can be explained by the fact that the SVM algorithm does not perform well with imbalanced data and the oversampling and undersampling procedures were unable to completely solve this predicament (still, much better results were obtained when employing these techniques than with the original data).

In spite of these results not being optimal, they could still be useful to predict bleeding events. Since similar percentage values were obtained for the sensitivity and specificity (in the order of the 75%), we could foresee with a reasonably high confidence both the occurrence or not of a bleeding for a certain patient. However, this would always lead to a misclassification error in the order of the 25% in both cases, which could be risky when such a serious event is at stake .

The fact that the dataset where the Principal Component Analysis was applied provided the best results for this algorithm is an illustration of how important removing superfluous information brought by the use of correlated variables in a model can be. To the Principal Components only the *Chadsvasc* feature had to be added to obtain the best results. This is proof of the high importance of this feature to explain the bleeding events.

D. Random Forests

The next method that was implemented to try to solve this problem was Random Forests. This method is based on decision trees which are a classification model that will try to classify each data point at each of the nodes(each node representing one variable) and check for information gain. It will then classify at the node where information gain is maximum. This process is repeated until there is no further information gain. Random Forests creates multiple decision trees and then combines the output generated by each one of them. While decision trees uses all the variables and observations, random forests will select a subset of observations and variables to build each tree. Each new observation will then be assigned to the mode of the classes of the individual trees.

Since the *train* function only tuned the parameter *mtry*, responsible for the number of variables that are available for the construction of each tree, and it was found that this parameter was not significant in obtaining better results it was decided to use the *randomForest* function. By using this function it was possible to find the optimal value for a much more relevant parameter to reduce the misclassification error, *ntree*, which gives the number of trees created. For each simulation, it was chosen the minimum value of *ntree* for which the out-of-bag error stabilized. The values for the *mtry* parameter and for the *sampsize* parameter (determines the number of observations used to construct each tree) were set as default for the simulation, since these were the ones which generated the best results.

For this method the best results were obtained for the model produced by the **data.trainUnder** considering the variables *Total Days*, *No. Days w/ INR<2*, *No. Days w/ INR>3*, *Age*, *Vascular*, *Chadsvasc*, *Age* $\geq$ *75* and *TTR<60%*. The second best results were generated by the **datapca.trainUnder** considering the four principal components and all the categorical variables. These results can be observed in the Table V.

	data.trainUnder w/ chosen variables	datapca.trainUnder
Accuracy	0.5312	0.5208
Sensitivity	1.00000	0.88889
Specificity	0.48276	0.48276
Youden's index	0.48276	0.37165
AUROC	0.7414	0.6858

confusion matrix

prediction \ reference	0	1
0	42	0
1	45	9

prediction \ reference	0	1
0	42	1
1	45	8

TABLE V: Random Forests Results

Since the Random Forests method is quite sensitive to class imbalance it was no surprise that the results obtained with the original dataset were so bad: a sensitivity of 0 and a specificity of 0.98851. These values are easily explained by the fact that the model learned to predict the majority class, ignoring the minority one. With this method adding more data, i.e. doing oversampling, didn't resolve the problem since the results obtained were almost the same as the ones obtained with the original dataset. The best results were all obtained with the datasets that had suffered undersampling. Even so, the results were very poor as shown in the Table V. Although the sensitivity is 1 (or almost 1) the values of the specificity are below 0.5, which leads to a high misclassification rate. These bad results could be explained by the fact that the model was trained using few samples. In general, the more imbalanced the dataset the more samples will be discarded when undersampling, therefore throwing away potentially useful information about the majority class. This could be one of the reasons why the models obtained can easily classify the bleeding episode, but not the non-bleeding one.

E. Association Rules Mining

As an alternative method to get more insights about the overall data patterns regarding the *Episode Bleeding* variable

association rules mining (ARM) was used. The goal of association rules mining is to find interesting association and interesting structures among the data. Although it is mostly used for transaction data in market and risk management, medical usage of the method is also rational. ARM discovers rules ($X \Rightarrow Y$) that indicate the likely occurrence of an item based on the occurrences of other items in the transaction. Let X be an itemset, $X \Rightarrow Y$ an association rule, t a transaction and D a transaction dataset [3]. Three main measurements are used to describe the association rules:

- Support: $supp(X) = \frac{|t \in T; X \subseteq t|}{|T|}$, i.e. the proportion of transactions t in the dataset which contains the itemset X
- Confidence: $conf(X \Rightarrow Y) = \frac{supp(X \bigcup Y)}{supp(X)}$, i.e. the proportion of the transactions that contains X which also contains Y. It describes how often the rule has been found to be true.
- Lift: $lift(X \Rightarrow Y) = \frac{supp(X \bigcup Y)}{supp(X) * supp(Y)}$, i.e. the ratio of the observed support to that expected if X and Y were independent. Lift being more than one indicates the degree those two occurrences' dependence on each another thus making the rules useful in predicting future outcomes.

The way ARM is applied to this dataset is the following; dataset **arm** contains all binary categorical variables of the pre-processed data and is being treated as a transaction data. As the variable of interest is *Episode Bleeding*, the rules having *Episode Bleeding* $= 1$ in the right hand side are the ones that give insightful information. The metric of interest is the lift as we are interested in which event occurrences accompany the occurrence of bleeding. Setting the threshold values of $conf = 0.08$ and $supp = 0.01$ the following patient conditions are among the rules having highest lift values.

- *Female*=0, *Age*$\geq$*75*=1, *Heart Failure*=1, *Hypertension*=1, *Diabetes*=1, *Stroke*=0 with lift of 8.18.
- *Female*=0, *Age*$\geq$*75*=1, *Heart Failure*=1, *Diabetes*=1, *Stroke*=0 with lift of 7.01.
- *Age*$\geq$*75*=1, *Hypertension*=1, *Diabetes*=1, *Stroke*=0, *Vascular*=1 with lift of 6.54.

Thus, for example, a male patient older than 75 and had a heart failure, hypertension, diabetes but no stroke, than it is 8.18 times likely for him to have bleeding. All other generated rules can be accessed in the R script of the project.

V. Discussion

The results in terms of the Youden's Index and AUROC for the 4 classification methods are included in the Table VI. It can be observed that the best results, by a large margin, for both of those measures were obtained with the KNN algorithm. The reason for such good results can be the similarity of medical records of the patients having bleeding episodes. Although such hypothesis should be confirmed by the medical specialists, the mathematical idea lying behind the algorithm points to such conclusions. The next best results were obtained for the Logistic Regression and the worst results were obtained for the Random Forest and SVM. KNN performed extremely well for this problem, as the Logistic Regression also achieved reasonable results, but they should have been better since this method is known for performing well with unbalanced data. The lack of descriminatory power of the features presented in the data to explain the bleeding events could be an explanation for this. As for the SVM and Random Forest algorithms, their results were not very satisfactory, which could be explained by the fact that both of these methods do not perform well with unbalanced response variables and the oversampling and undersampling techniques were unable to completely overcome this issue, although they improved, by far, the obtained results.

	Dataset which gave the best results	Youden's Index	AUROC	Variables which gave the best result
Logistic Regression	data.train	0.6169	0.8282	*Total Days* *Total No. of Tests* *No. Days w/ INR>3.5* *TTR<60%*, *Chadsvasc* *(Total No. of Tests* *Total Days)*
KNN	data.trainSMOTE	0.7087	0.9719	*Total Days* *Total No. of Tests* *Age*, *Age≥75* *Vascular* *Chadsvasc*
SVM	datapca.trainBoth	0.5364	0.7682	4 princ. components + *Chadsvasc*
Random Forests	data.trainUnder	0.4828	0.7414	*Total Days* *No. Days w/ INR>3* *No. Days w/ INR<2* *TTR<60%* *Age*, *Vascular* *Chadsvasc* *Age≥75*

TABLE VI: Comparison of the best models of each method

In all of the models, the best results were obtained for a reasonably low amount of variables (maximum of 8 covariates). Many of these features were repeatedly selected for the different models. The *Chadsvasc* is the only feature included in all of the best models. The variables *Total Days*, *Total No. of Tests*, *Age*, *Age≥75*, *TTR<60%*, *Vascular* appear in at least 2 of the models. Hence, these variables can be considered as the most statistically important for this problem.

It is also curious to notice that the datasets for which the best results were obtained in each method were different. Thus, there is not a general optimal dataset that works for every method, which is an illustration of how important it is to try several data pre-processing methods to generate the best results.

VI. Conclusion

In conclusion, four supervised learning algorithms were applied to VKA treatment patients' medical dataset along with an additional Association Rules Mining technique. As a result of exploratory data analysis and the applications of the algorithms, the following variables were found to have significant importance on the event of bleeding: *Total Days, Total Number of Tests, Age, Vascular, Chadsvasc, No. Days*

w/ INR<2, No. Days w/ INR>3.5, TTR<60. It comes as no surprise that the *Total Days* and *Total Number of Tests* were significant to our models, since, the longer a patient remains in the study, obviously the higher the chances that he/she will eventually suffer a clinical adverse events, whichever they may be. *Age* is also obviously connected with the occurrence of an adverse event, as elderly people are more vulnerable to these kind of episodes. The significance of the *No. Days w/ INR>3.5* feature is also easily explained by the fact that higher values of INR are connected with a higher fluidity of the blood, which provokes the bleedings. An interesting insight is the importance of the variable *No. Days w/ INR<2* on the bleeding event; *INR<2* indicates thickness of blood whereas bleeding is caused in case of higher INR, i.e. more fluid blood. Thus, the variable helps to indicate non-bleeding events. *Chadsvasc*, on account of containing information on multiple variables related to the risk of thromboembolic events was constantly selected as one of the most important features for all models. This is due to the fact that a thromboembolic event is connected to low values of INR, consequently allowing us to discriminate a non-bleeding event. From the features used to calculate the *Chadsvasc*, only *Vascular* was found to be relevant for this classification problem. Even though this variable is not related to the thickness of the blood, it characterizes the condition of the blood vessels. As for the *TTR<60%*, even though this variable was only found to be meaningful for two of the four applied supervised classification methods, it shouldn't be overlooked since it's helpful to understand if there was a good anticoagulation control, or not, for that patient. However, all of these previous medical inferences about these features could not be valid for all of the methods and models tested. As it was observed for the case of the Logistic Regression with the Odds Ratio for the *Chadsvasc* and the *No. Days w/ INR>3.5*, different mathematical interpretations could be reached and, even though, they differ from an initial medical intuition of the problem, they are still essential to obtain good results. It was surprising to find out that the Gender does not have any effect on the bleeding variable as there are many physiological differences between males and females that might affect the bleeding. The results of ARM can also be of great medical help as having observed a certain type of health behavior, the medical team can expect a certain likeliness of confronting bleeding.

Not being able to achieve high accuracy measures in the classification methods employed, except for the KNN, may be explained by the low sample size and unbalanced response variable or we could be dealing with the case where the features have a poor discriminative power, and therefore adding more data of the same type won't necessary help to improve the results. Having many values that were generated through interpolation could also have affected negatively the outcomes. Since, we only know for sure the values related to the INR for a certain patient when they are tested (time between tests can vary between a few days to a few months), misleading information about the values of INR a patient has in-between tests could be generated, which influenced the results. However, to make an analysis of this type, the exclusive usage of the values of the INR during the tests is clearly insufficient. Hence, the easiest way to reduce the uncertainty related with this type of approach is to increase the number of tests applied to each patient, even if that patient has seemingly achieved a balanced INR state. The high number of correlated variables in the dataset could also have hindered the attainment of better results. Even though we were dealing with a large number of features, as most of them present high correlations with each other, the amount of information that is not redundant is very slight. Therefore, the inclusion of some other relevant variables for this analysis, independent from the ones already included, could lead to a vast improvement of the results obtained.

One of our suggestions for further research and improvement is obtaining required data for HAS-BLED scoring system which has a great predictive value in bleeding events and consists of the following features: *Hypertension, Abnormal renal function, Abnormal liver function, Stroke, Bleeding, INR, Elderly: Age > 65 years, Prior Alcohol or Drug Usage History , Medication Usage Predisposing to Bleeding*. It is also surprising that variables *Stroke* and *Hypertension* did not appear among the important variables throughout the analysis as they are included in the HAS-BLED scoring system and should have an impact on bleeding. Although the majority of the features required by HAS-BLED were available in the data, some of the features are missing. We believe that having these features would significantly improve the results.[4]

REFERENCES

[1] Chawla N., Bowyer K., Hall L., Kegelmeyer P. *SMOTE: Synthetic Minority Over-sampling Technique*,Journal of Artificial Intelligence Research 16 (2002) 321–357
Retrieved from: https://arxiv.org/pdf/1106.1813.pdf

[2] Li-Yu Hu, Min-Wei Huang, Shih-Wen Ke, Chih-Fong Tsai *The distance function effect on k-nearest neighbor classification for medical datasets*, Hu et al. SpringerPlus (2016) 5:1304
Retrieved from: https://pdfs.semanticscholar.org/d258/b75043376e1809bda00487023c4025e1c7a9.pdf

[3] Kotsiantis S., Kanellopoulos D. *Association Rules Mining: A Recent Overview*, GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82
Retrieved from: http://www.csis.pace.edu/~ctappert/dps/d861-13/session2-p1.pdf

[4] Lane D.A., Lip G.Y.H. *Use of the CHA2DS2-VASc and HAS-BLED Scores to Aid Decision Making for Thromboprophylaxis in Nonvalvular Atrial Fibrillation*
Retrieved from: http://circ.ahajournals.org/content/126/7/860