DATA MINING, MACHINE LEARNING, AND DEEP LEARNING
CDSCO1004U.LAF21

Copenhagen Business School

Final Project Group Fri-103831-17
(Research Paper)

Submitted By

Helena van Eek (141521)
Magnus Beck Eliassen (141855)
Sabrina Breunig (141706)
Eirik Egge (141164)

Master of Science in
Business Administration and Data Science

Lecturer
Raghava Rao Mukkamala
Associate Professor, PhD

Number of pages: 15
Number of characters: 33.886

26 May 2021

Supervised Machine Learning Approach for the Classification of Stroke Risk based on Imbalanced Medical Data

Abstract

Due to stroke being the second leading cause of preventable death worldwide, it is vital to understand the medical conditions and lifestyle factors impacting the risk of the disease to conduct medical examinations promptly. Consequently, data-driven solutions for accurately identifying a patient's stroke risk are valuable for heightening chances of prevention and early treatment. For this purpose, this paper presents supervised machine learning techniques leveraging medical data for the prediction of stroke risk. The problem investigated revolves around the vast imbalance of the examined medical dataset, causing the algorithms' failure to detect a single stroke risk in the baseline models. To master this challenge, this paper first applies the oversampling techniques Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN) to then finally tune the classifiers Support Vector Machines (SVM), Random Forest (RF), Multi-Layer Perceptron (MLP), and the majority vote on the synthetically balanced dataset to boost the recall score of the minority class as the prioritized performance heuristic. The results highlight SVM combined with SMOTE as the favored classifier due to its recall score of 1.0 for correctly predicting "stroke" with a false positive rate of 60%. The presentation of the variables heart disease, hypertension, age, residence type, and average glucose level as the most important features for the proposed classifier finalizes the conducted analysis.

Keywords: *Stroke prediction, Imbalanced data, SMOTE, ADASYN, SVM, RF, MLP, Majority vote*

1 INTRODUCTION

Since the outbreak of the COVID-19 pandemic, the threat of stroke as the globally second leading cause for preventable death has further intensified (Kuklina et al., 2014, p. 1). Worldwide lockdowns resulted in increased burdens for parents, anxiety caused by social distancing, and a profound lack of physical activity (Prati Mancini, 2021, p. 203f). Ultimately, all these factors impacting physical and mental health combined result in an amplified risk of stroke (Kuklina et al., 2014, p. 1). In addition, the demand for hospital beds for COVID-19 patients coupled with the fear of infection prevents several stroke patients from undergoing medical examinations in times of the pandemic, reducing their chances for prevention and early treatment (Dula et al., 2020, p.1). Hence, to address the global challenge of stroke, supervised machine learning techniques offer the potential to unfold the power of medical data to enhance patient care by classifying individual health records into predefined diagnosis labels (Dash et al., 2019, p. 2-3). However, manifold limitations of classification in healthcare arise as medical datasets are often highly imbalanced caused by the fact that only a minority of the population suffers from a particular disease (Khalilia et al., 2011, p. 5). Consequently, predictions may lack desirable performance, resulting in even useless classifications in the worst case. Thus, taking into account the challenges of the global health situation from the medical perspective and of imbalanced data from the technical perspective, the following research question arises:

How and to what degree does imbalanced medical data allow the prediction of stroke risk to assist its prevention and early treatment through supervised machine learning, and which medical patient features are the most important for doing so?

To answer the above, the focus of this paper lies in the following five propositions and their confirmation or rejection by the presented analysis.

Proposition 1: Accuracy is a misleading performance metric for the given dataset as classifiers tend to always predict the majority class "no stroke" due to the severely skewed data.

Proposition 2: Balancing the dataset with SMOTE and ADASYN significantly improves precision and recall of the baseline models for correctly classifying patients with stroke to a degree that renders the classifier valuable as a real-world application.

Proposition 3: The hyperparameter tuning of the classifiers after balancing highlights the immense trade-off between precision and recall for highly imbalanced data that must be resolved with respect to a particular medical real-world application.

Proposition 4: The ensemble technique majority voting outperforms the three individual classifiers SVM, RF, and MLP in terms of precision and recall scores.

Proposition 5: The five most important features for stroke prediction are a patient's age, BMI, smoking habits, heart diseases, and hypertension.

Focusing on these propositions the paper is structured as follows. In Section 2 related work investigating machine learning in healthcare is presented. Next, the dataset description and the methodology for the applied processes and methods are found in Section 3. Further, the core findings of the conducted data analysis including performance metrics and model complexity are presented in Section 4, followed by a discussion and limitations in Section 5. Finally, the paper presents the conclusion in Section 6.

2 RELATED WORK

A variety of existing literature examines stroke prediction through supervised machine learning leveraging balancing techniques and classification algorithms. In light of the medical perspective, Dash et al. provide a general overview of the implications, opportunities, and challenges of machine learning in the medical sector (Dash et al., 2019). Ellis et al. elaborate on common causes of a stroke to increase symptom awareness for the prevention of the disease (Ellis et al., 2013). In light of the technical perspective, Islam highlights the use of Fuzzy logic Inference Systems and C-means fuzzy classifier to predict the deadly disease (Islam, 2018). Khosla et al. compare the performance of a margin-based censored regression algorithm with the Cox proportional hazards model on the Cardiovascular Health Study dataset with a special focus on feature selection (Khosla et al., 2010). Further, Hamed et al. investigate the performance of SVM and Neural Networks using SPSS®, MATLAB®, and Rapidminer® (Hamed et al., 2014). A hybrid method is presented by Liu et al. who combine missing value imputation and an AutoHPO-based DNN prediction model for the supervised prediction for stroke risk (Liu et al., 2019). Moreover, Belarouci Chikh balance the medical data with a cost-sensitive extension of the Least Mean Square algorithm for the following performance comparison of SVM, K-Nearest Neighbors, and Multilayer Perceptron techniques (Belarouci Chikh, 2017). Lastly, Khalilia et al. address the challenge of a highly imbalanced dataset by balancing the data with repeated random subsampling to then predict the risk of diseases with RFs (Khalilia et al., 2011). The approach presented in this paper differs from the above by applying the oversampling methods SMOTE and ADASYN to synthetically achieve class balance for the ultimate supervised prediction of stroke risk through the classifiers SVM, RF, and MLP that are tuned in light of the recall score for the minority class.

3 CONCEPTUAL FRAMEWORK

An imbalanced dataset is characterized by an unequal distribution of the classes in the training dataset (Ganganwar, 2012, p. 1). This property is a crucial concern for real-world applications where misclassifying an observation of the highly underrepresented class is associated with high costs, i.e. in the case of disease detection (Chawala, 2010, p. 1). The imbalance poses profound challenges on classification algorithms designed for an equally distributed dataset (Rahman & Davis, 2013, p. 227). As a consequence, machine learning techniques perform extraordinarily well on the majority class but often fully fail to detect the minority class and hence, result in classifiers without any use (Ganganwar, 2012, p. 1). Thus, in light of performance heuristics, recall is the most appropriate choice in use cases with high costs associated with false negatives and hence the focus of the performance comparisons in this paper (Chawla et al., 2002, p. 322, Géron, 2017, p. 124). To optimize the recall scores for the minority class, the f-beta scoring that favors recall through a beta = 2 is applied for the hyperparameter tuning.

Further, balancing techniques such as oversampling and undersampling address the problem of the classifier learning bias towards the majority class (Chawla et al., 2002, p. 321). While oversampling replicates data points from the minority class to increase its cardinality, undersampling samples the majority class to reduce its data points (He et al., 2008, p. 1322). For the purpose of this paper, the two oversampling

techniques SMOTE and ADASYN are introduced to prevent the loss of a large amount of valuable data.

4 RESEARCH METHODOLOGY

4.1 Dataset Description

The raw dataset[1] is obtained from Kaggle, published by Frederico Soriano Palacios. In total, the dataset consists of 5110 rows, each describing an individual patient. Further, the 12 different features are separated into independent variables and the target variable. Prior include the numerical variables 'Id', 'age', 'hypertension', 'heart_disease', 'avg_glucose_level' and 'bmi' additionally the categorical variables 'gender', 'ever_married', 'work_type', 'Residence_type', and 'smoking_status'. The binary target variable presenting 'no stroke' or 'stroke' for each patient shows an imbalanced ratio of 96,4% to 3,6%, respectively. Examining the completeness of the data, only the variable 'bmi' includes missing values, accounting for 4% of the samples.

4.2 Data Preprocessing & Analysis Process

The data analysis process is illustrated in Figure 1. The machine learning methods are performed using the integrated functions of the scikit-learn package in Python.

4.2.1 One Hot Encoding & Correlation Examination: Categorical data requires a numerical transformation to be interpreted by machine learning models. Categorical variables with just two values were binarized. For all other categorical variables, we apply One Hot Encoding; hence one binary attribute is created for each value. Contrary to other common encoding techniques, One Hot Encoding does not imply an order between the given nominal features. (Géron, 2017, p. 92)
After the encoding the correlation matrix is examined. As none of the correlation coefficients are higher than 0.7 we confirm that there is no alarming high correlation between two independent variables (Ratner, 2009, p.140). In addition, the existence of multicollinearity is analyzed by calculating the Variance Inflation Factor (VIF) scores for each variable. As high correlation is considered for values over 10 the feature called 'private job' with a VIF score of over 20 is eliminated from the dataset. Thereby, multicollinearity does not undermine the interpretation of the models. (Montgomery, 2001, p. 672)

Figure 1: Data Preprocessing and Analysis

4.2.2 Outliers: An outlier is a data object that deviates significantly from the rest of the objects as if it were generated by a different mechanism (Han, Kamber, & Pei, 2012, p.544). In our dataset, there are abnormal values for 'bmi' and 'avg_glucose_level'. Whether it is advantageous to remove them varies with their importance, but such importance can be difficult to determine. (Han, Kamber, & Pei, 2012, p.22). For medical data, it is essential not to remove data that reduces the exactness of the result. Health data may have outliers for many different reasons, one of which is abnormal patient conditions (Singh & Upadhyaya, 2012, p. 315). Therefore, outliers may be highly relevant to the prediction of diseases. Considering our dataset, removing too many outliers may cause our models to predict false negatives where stroke is likely, as we would have less data on patients with abnormal BMI or glucose level. However, these patients could be the ones at greater risk of stroke. Another factor that reassures this decision is

1: Visit https://www.kaggle.com/fedesoriano/stroke-prediction-dataset to view the original dataset.

the number of abnormal values in the dataset. The features 'bmi' and 'avg_glucose_level' both contain many relative to the size of the dataset. If we use IQR scores to define the middle 50% of the data as a basis for detecting and removing outliers, it will remove close to 20% of the entire dataset. Considering the size of the dataset and the potential impact of the abnormal data on the results, this is not desirable. Thus, we only remove the most extreme outliers to not bias our results. Instead of calculating the IQR-score as the difference between the 75th and the 25th quartile, we use the 90th and the 10th quantile. Thus, the rows with the most extreme abnormal data from the 'bmi' column are removed, resulting in four fewer rows. Since we retain such a large proportion of the dataset, we also conducted a parallel model removing a greater proportion of the outliers to ensure reliable results, which can be found in appendix Section 8.2. Thereby, we confirmed our prior decision.

4.2.3 Feature Scaling: In case the numerical attributes have considerably different scales, most machine learning algorithms do not perform well. To prevent this, two common scaling techniques are normalization and standardization, which both have subsets of various approaches. By normalizing features, their values are rescaled to a certain range, typically from zero to one (Géron, 2017, p.95). A common normalization technique is Min-Max scaling which preserves the shape of the original distribution. However, as two of the base terms in the expression are the minimum and maximum value of each feature, this technique is highly sensitive to outliers. Furthermore, by standardizing features, each feature achieves a zero-mean and the distribution is scaled to unit variance. Contrary to normalization techniques, standardization does not bound values to a certain range in addition to being far less affected by outliers. (Charu C. Aggarwal, 2018, p. 147) The performance of the different machine learning models might vary regarding which scaling technique is used. For instance, artificial neural networks often expect input values ranging from zero to one (Géron, 2017, p.95). In appendix Section 8.1, model performance measures are attached for the different scaling techniques. The performance measures vary across different models; consequently, this project has applied Min-Max scaling for MLP, and standard scaling for RF and SVM.

4.3 Transformation of Imbalanced Data

4.3.1 SMOTE: SMOTE replicates observations in the minority class by identifying the k-nearest neighbors for each sample x e A. It does so by first calculating the Euclidean distance between every x and all other observations in A. Next, a sampling rate N is defined according to the imbalanced proportion so that for every x e A, N points are randomly selected from the k-nearest neighbors of x to generate a new subset of A, named A_1. Afterward, the algorithm creates new samples for every x_k e A_1 along the following equation whereby rand (0,1) generates a random number between 0 and 1:

$$x_{new} = x + rand\,(0,1) * |x - x_k| \tag{1}$$

Finally, by adding these newly generated samples to the minority class a new dataset is created. (Chawla et al., 2002, p. 328-329)

4.3.2 ADASYN: As an extension of the SMOTE technique, ADASYN attempts to balance the underlying dataset using linear interpolation to synthetically generate new observations for the underrepresented class A. To do so, first, the number of newly generated observations G is calculated by the formula:

$$G = (m_r - m_x) * \beta \tag{2}$$

whereby m_r and m_x represent the number of data points in the minority and majority class, respectively, and β refers to the balance level of the newly generated samples. In case of β = 1, both classes would be equally balanced. Similar to SMOTE, for each minority datapoint x_i e A, the Euclidean distance is used to identify the k-nearest neighbors. In contrast to SMOTE, next the ratio r_i is determined for each x_i to then calculate the density distribution r_x along the following equation:

$$r_x = \frac{r_i}{\sum r_x} \tag{3}$$

In the next step, the number of synthetic samples g_i to generate for each $x_i$are defined using the density distribution r_x according to the following formula:

$$g_i = r_x * G \tag{4}$$

Finally, g_i observations for each x_i are created along

the following steps within a loop from 1 to g_i. First, one minority data sample, called x_u, is randomly selected from the k-nearest neighbors of x_i. Secondly, for each x_u, the synthetic data point s_i is calculated with the following equation whereby rand(0,1) generates a random number between 0 and 1:

$$s_i = x_i + (x_u - x_i) * rand\,(0, 1) \tag{5}$$

Lastly, a new dataset is created by adding all s_i to the minority class. Thus, ADASYN presents an adaptive learning procedure that dynamically adjusts the ratio r in contrast to SMOTE which uses a uniform weight for all minority points. Consequently, while SMOTE creates synthetically samples in the interior of the minority class, ADASYN generates observations in the vicinity of the boundary between the minority and majority class. (He et al., 2008, p. 1323f)

4.4 Supervised Classification Algorithms

4.4.1 Support Vector Machines: SVM is a powerful and versatile machine learning model for both linear and nonlinear classification, regression, and detection of outliers. SVM is commonly used to classify relatively complex datasets of smaller sizes. As our dataset fits this description given the number of samples and the high imbalance of the target variable, we believe that SVM is highly capable of distinguishing potential strokes from no stroke compared to other classifiers. SVM uses support vectors to find the best possible hyperplane between different groups of data. Thereby, the model manages to fern the hyperplane with the most significant margins to the various data points. The more considerable such a margin, the better the model will classify future data (Han, Kamber, & Pei, 2012, p. 409). SVM is also beneficial as, unlike MLP, it always finds a global solution (Han, Kamber, Pei, 2012, p. 415). For our dataset, we test out different kernels by fine-tuning our parameters utilizing grid search. Based on the chosen kernel, the training complexity lies between $O(m^2 \cdot n)$ and $O(m^3 \cdot n)$ (Géron, 2017, p. 211). The other parameters tuned with grid search are 'C', 'gamma', 'degree', and 'max_iter'. (Géron, 2017, p. 201)

4.4.2 Random Forest: RFs are a powerful and widely used decision tree classification model, which gathers a collection of classifiers in a "forest". For classification, every "tree" votes and the class with the most votes will be returned. There are multiple advantages associated with RFs. For instance, they swap higher bias with a lower variance by searching for the best features in a random subset of features, which leads to a more solid performance (Géron, 2017, p. 255). It also handles outliers and errors well compared to other classification models. For our dataset, this is advantageous given the prior introduced outlier removal approach. Also, overfitting is no highly relevant concern since the generalization error can be reduced with a high number of trees. For RFs, each classifier needs to maintain its strength without increasing its interdependency to return high accuracy. Hence, RFs are sensitive to the number of attributes taken into account at each split, where few attributes and, in some cases, just one attribute can create excellent accuracy. (Han, Kamber, & Pei, 2012, p. 383) Defining s as sampling batches, c as the average decision tree node number, m as the feature number of each tree, and n as the average length of sampling bathes, the approximate computational complexity of ensemble random forests equals up to $O(s \cdot c \cdot m \cdot n \cdot log(n))$ (Géron, 2017, p. 254f). To improve the performance of the model, grid search is performed for tuning the hyperparameters 'min_samples_split', 'n_estimators', 'max_depth', 'max_features', and 'ccp_alpha'.

4.4.3 Multi-Layer Perceptron: An MLP is an artificial neural network consisting of one input layer and one or more hidden layers. In cases with more than one hidden layer, it is referred to as a deep neural network. MLPs are typically used for binary classification of imbalanced datasets. In MLPs, the neurons calculate the weighted sum of their inputs. Furthermore, an activation function is applied to acquire a signal that is transmitted to the next neuron (Castro et. al, 2017). The final output of each neuron is the estimated probability of the corresponding class. In order to train MLPs, the backpropagation training algorithm is used. It first applies a forward pass and measures the error. Further, it reverse-passes through each layer and measures the contributed error from each connection. Finally, the connection weights are tweaked leading to a reduction in error. For MLPs, the application of nonlinear activation functions is important. Common activation functions for MLPs are sigmoid, hyperbolic tangents, and rectified linear unit (ReLU) (Aggarwal, 2018, p.11-14). The computational complexity of MLPs highly depends on

the underlying architecture of the model (Aggarwal, 2018, p. 35f). To decide on the most suitable architecture for the purpose of this paper, hyperparameter tuning for 'activation', 'hidden layer sizes', 'solver', 'alpha', and 'learning rate' is applied. (Géron, 2017, p.350-352)

4.4.4 Voting Classifier: The voting classifier is an ensemble model that combines the predictions of multiple classifiers to resolve one model's error with another model. For hard voting, the predicted classifications of the voting classifier are calculated as the mode of the distribution of the predictions given by the individual classifiers. (Géron, 2017, p. 246)

4.4.5 Feature Importance: After performing the classification it is of interest to investigate the importance of features for predicting the target variable. This technique provides insights into the model and the data by allowing an understanding between the input features and the output. Each input variable gets assigned a score based on different types of calculation methods depending on the classifier. For RF, the relative feature importance is based on impurity-based feature importances and can be accessed after the model is fitted. This is performed by calculating the decrease in the impurity of the split for each feature and taking the average over all trees in the forest. For binary classification, the so-called "Gini importance" is calculated. An advantage of this method is the computation of all needed values as part of the RF training leading to a fast computation. (Saarela Jauhiainen, 2021, p. 4)
Further, the most important features of MLPs are calculated by using the permutation importance technique. While performing permutation the relative feature importance scores are calculated independently from the model used. To do so, single input variables are randomly shuffled while preserving the others and the effect on the final prediction performance is measured. By corrupting the structure of the data for the single variable, the corresponding relationship is broken. If this results in a higher error, this variable is of high importance for the prediction. (Varma K., 2020, p.2)
Lastly, for linear SVM, feature importance can be retrieved from the coefficients of the trained model. The coefficients are the weights representing the coordinates of a vector that is orthogonal to the hyperplane and provide a raw score for feature importance. The comparison of the size of weights offers clarity as features holding less variance are less important for the prediction. (Weston et al., 2000, p.45)

5 FINDINGS

This section is dedicated to the results of our data analysis and the examination of whether the five propositions are to be confirmed or rejected.

5.1 Proposition 1

Accuracy is a misleading performance metric for the given dataset as classifiers tend to always predict the majority class "no stroke" due to the severely skewed data. After applying the baseline classifiers, the classification reports of all three individual classifiers show an accuracy of 94%. To give an illustration, Table 1 shows the classification report for RF. Further investigations highlight that always predicting the majority class results in an accuracy of 94% only by neglecting the minority class completely. Hence, as SVM and MLP show very similar results, these classifiers are without use in an application that strives to predict a high number of samples in the minority class correctly. Consequently, accuracy should only be used very cautiously alongside other heuristics in case of imbalanced datasets. As false negatives have the highest cost associated within the given context, recall is the focus heuristic of the following model evaluations.

Table 1: Classification Report RF Baseline Model

	Precision	Recall	F1-Score	Support
0	0.94	1.00	0.97	1197
1	0.00	0.00	0.00	80
accuracy			0.94	1277
macro avg	0.47	0.50	0.48	1277
weighted avg	0.88	0.94	0.91	1277

5.2 Proposition 2

Balancing the dataset with SMOTE and ADASYN significantly improves the precision and recall of the baseline models for correctly classifying patients with stroke to a degree that renders the classifier valuable as a real-world application. As SMOTE and ADASYN require a running time of 0.01 and 0.02 seconds, respectively, the model complexity is considered neglectable for comparing the two balancing methods. In light of

performance heuristics, the three classifiers show the results presented in Table 2 and Table 3 for the balanced training data.

Table 2: Precision & Recall SMOTE w/o Tuning

	SVM		RF		MLP	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
No Stroke	0.98	0.70	0.95	0.91	0.98	0.72
Stroke	0.14	0.75	0.18	0.30	0.15	0.74

Table 3: Precision & Recall ADASYN w/o Tuning

	SVM		RF		MLP	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
No Stroke	0.98	0.67	0.95	0.90	0.98	0.71
Stroke	0.14	0.79	0.14	0.25	0.15	0.78

In terms of recall for predicting "stroke", ADASYN outperforms SMOTE for SVM and MLP. Overall SVM with ADASYN performs the best at this point of the analysis, achieving a recall of 0.79. In light of the results above, the second proposition is confirmed in view of the authors.

5.3 Proposition 3

The hyperparameter tuning of the classifiers after balancing highlights the immense trade-off between precision and recall for highly imbalanced data that must be resolved with respect to a particular medical real-world application. Tables 4 and 5 show the performance of the classifiers after tuning the hyperparameters for the data balanced with SMOTE and ADASYN. Taking into account all results, the trade-off between precision and recall is clearly highlighted. As this analysis focuses on the recall for predicting “stroke”, SVM with SMOTE is considered the best performing model for the given dataset. Considering the fact that all patients with a stroke risk are classified accordingly, the false positive rate of 60% is valued as acceptable in this specific case. Further, in light of model complexity measured in running time, both versions of SVM and RF need less than 1 second for completion. Only the MLP classifier with SMOTE and ADASYN require a significantly higher running time of 20.98 and 44.97 seconds, respectively, given the models’ complexity. Compared to the baseline model of SVM with an accuracy of 94%, the tuned SVM classifier with the SMOTE balanced dataset achieves an accuracy of only 44%. However, given the fact that the recall for predicting stroke increased from 0.0 to 1.0, the decrease in accuracy is found tolerable.

Table 4: Precision & Recall SMOTE w/ Tuning

	SVM		RF		MLP	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
No Stroke	1.00	0.40	0.98	0.65	0.97	0.75
Stroke	0.10	1.00	0.14	0.84	0.15	0.64

Table 5: Precision & Recall ADASYN w/ Tuning

	SVM		RF		MLP	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
No Stroke	0.00	0.00	0.98	0.65	0.98	0.71
Stroke	0.06	1.00	0.14	0.84	0.15	0.78

In contrast, the SVM classifier with the ADASYN dataset also achieves a recall for predicting stroke of 1.0 but has a false positive rate of 100%, demonstrating the case that despite the recall of the minority class being the prioritized performance heuristic, false positive rates should not be ignored to ensure valuable results. Otherwise, all patients are classified with “stroke”, resulting in a classifier without any use in the real world. Further, in case a false positive rate of 60% is perceived as unacceptable in a specific real-world application, the ROC curve in Figure 2 is useful for further inspection. It highlights the trade-off between true positive rates and false positive rates for all classifiers. Thus, the ROC curve allows identifying the most suitable classifier satisfying the desired trade-off given a particular application.

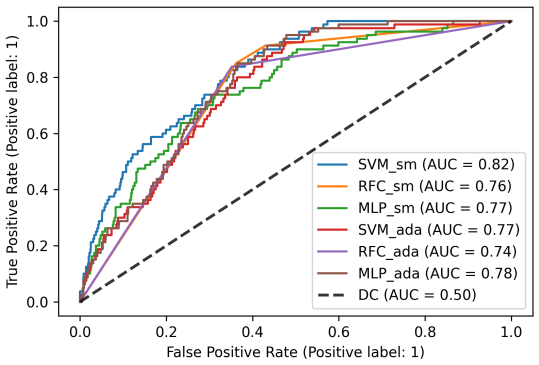


Figure 2: ROC-Curve for all tuned classifiers

5.4 Proposition 4

The ensemble technique majority voting outperforms the three individual classifiers SVM, RF, and MLP in terms of precision and recall. Table 6 presents the comparison of the proposed SVM classifier to the hard voting classifier. Investigating the results, SVM clearly outperforms the ensemble method in terms of recall for "stroke". Also, the voting classifiers' running time with SMOTE and ADASYN accounts for 18.51 and 0.72 seconds, respectively, compared to 0.27 seconds for SVM. Consequently, proposition 4 is rejected and SVM combined with SMOTE remains the favored classifier.

Table 6: Comparison of SVM & Hard Voting

	SVM SMOTE		Voting Clf. SMOTE		Voting Clf. ADASYN	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
No Stroke	1.00	0.40	0.99	0.63	0.99	0.52
Stroke	0.1	1.0	0.13	0.86	0.12	0.94

5.5 Proposition 5

The five most important features for the best performing stroke prediction model are a patient's age, BMI, smoking habits, heart diseases, and hypertension. First, it is important to note that the different classifiers disagree in the order of the most important features and their respective scores which is further illustrated in Section 8.3 of the appendix. Table 7 shows the five most important features for predicting stroke using the recommended SVM classifier combined with SMOTE. In contrast to the expected most important features, the proposed classifier does not list BMI and smoking habits. Instead, residence type and glucose level contribute the most to successfully predicting stroke together with the expected features 'heart_disease', 'hypertension', and 'age'. Hence, proposition 5 cannot be confirmed.

Table 7: Most important features SVM

	Features	Coefficient
1	heart_disease	0.26
2	hypertension	0.17
3	age	0.09
4	residence_type	0.08
5	avg_glucose_level	0.07

6 DISCUSSION

6.1 Reflection on Research Question

The key aspiration of this study is to provide clinicians a predictive tool that adds value to their everyday work through clear explanations for fast and accurate stroke risk predictions. Hence, the outcome of our machine learning approach must be easily interpretable to provide meaningful facts and actionable insights for the investigation of patients' stroke risk. Based on the research question and findings presented in this paper, we believe to have successfully created an impactful predictive model with high scores for the prioritized performance heuristics, acceptable model complexity and profound risk factor interpretability. Also, our model significantly reduced the false negative rate compared to other traditional approaches for stroke risk prediction (Liu et al., 2019, p. 9). Moreover, the conducted analysis demonstrates that balancing the data with SMOTE and ADASYN profoundly enhanced the recall of the models while significantly increasing the false positive rates. Related work highlights that despite SMOTE and ADASYN being widely applied balancing methods, they may not be very beneficial if certain assumptions do not hold. For instance, according to Blagus Lusa, oversampling methods create a correlation between the samples, resulting in challenges for classifiers assuming independence among observations (Blagus Lusa, 2013, p.3). They highlight that even though oversampling improves results in comparison to the uncorrected data, the effect on the overall performance metrics can still be far from the one desired (Blagus Lusa, 2013, p.4). Nevertheless, coupled with hyperparameter tuning based on an f-beta scoring favoring recall, the classification of the balanced datasets achieves satisfactory results. Further, in light of feature importance, there is a clear overlap between the most important features identified in this analysis and the ones presented by medical research. This further verifies the proposed algorithm as suitable for stroke risk classification in a real-world application.

6.2 Limitations

Finally, we recognize the limitations of the presented analysis. First, information about the required effort resulting from a false positive would allow a well-founded decision for the trade-off between precision

and recall. This effort could be estimated using additional information regarding healthcare systems and living conditions. Besides, features such as stroke history of family members, alcohol drinking behavior, and amount of exercise would enable further improvements of the prediction. Future work should gather more enriched data and combine oversampling with undersampling methods to examine the effects on the predictive models.

7 CONCLUSION

This paper proposed a novel approach for predicting stroke risk to assist its prevention and early treatment in light of a highly imbalanced medical dataset. We demonstrated that applying oversampling methods and tuning the hyperparameters focusing on the recall for "stroke" tremendously improves the desired classification performance. Ultimately, using SVM with SMOTE balanced data achieves a recall for "stroke" of 1.0 and a false positive rate of 60% and hence, it is the proposed classifier for real-world applications based on our analysis. Further, in case this high false positive rate is not acceptable in a real-world setting, this paper also presented models achieving more balanced precision and recall scores, which can be chosen depending on the particular context. Lastly, the proposed classifier selected heart disease, hypertension, age, residence type, and average glucose level as the five most important features.

REFERENCES

Aggarwal, C. C. (2018). Neural Networks and Deep Learning. Springer International Publishing AG.

Belarouci, S., & Chikh, M. A. (2017). Medical imbalanced data classification. Advances in Science Technology and Engineering Systems Journal.

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. Journal of Artificial Intelligence Research.

Castro, W., Oblitas, J., Santa-Cruz, R., & Avila-George, H. (2017). Multilayer perceptron architecture optimization using parallel computing techniques. PLoS ONE, 12.

Chawala, N. V. (2010). Data Mining for Imbalanced Datasets: An Overview. Springer International Publishing AG.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, P. W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research.

Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: Management, analysis and future prospects. Journal of Big Data

Dula, A. N., Brown, G. G., Aggarwal, A., & Clark, K. L. (2020). Decrease in Stroke Diagnoses During the COVID-19 Pandemic: Where Did All Our Stroke Patients Go? JMIR Aging.

Ellis, C., Barley, J., & Grubaugh, A. (2013). Post-stroke Knowledge and Symptom Awareness: A Global Issue for Secondary Stroke Prevention. Cerebrovasc Dis.

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering.

Géron, A. (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow. O'Reilly Media, Inc.

Hamed, A., Dowling, R., Yan, B., & Mitchell, P. (2014). Machine Learning for Outcome Prediction of Acute Ischemic Stroke Post Intra-Arterial Therapy. PLoS ONE 9(2).

Hastie, T., Tibshirani, R., & Friedmann, J. (2001). The Elements of Statistical Learning. Springer Series in Statistics

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. Islam, F. (2018). A fuzzy logic based predictive model for early detection of stroke. International Joint Conference on Neural Networks.

Jiawei Han, Micheline Kamber, & Jian Pei. (2012). Data Mining Concepts and Techniques. Morgan Kaufmann Publishers is an imprint of Elsevier.

Johnson, W., Onuma, O., Owolabi, M., & Sachdev, S. (2016). Stroke: A global response is needed. Bull World Health Organ.

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak.

Khosla, A., Cao, Y., Chiung-Yu Lin, C., Chiu, H.-K., Hu, J., & Lee, H. (2010). An integrated machine learning approach to stroke prediction. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Kuklina, E. V., Tong, X., George, M. G., & Bansil, P. (2014). Epidemiology and prevention of stroke: A worldwide perspective. Expert Rev Neurother.

Liu, T., Wenhui, F., & Wu, C. (2019). A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. Artif Intell Med.

Montgomery, D. C., & Vining, G. G. (2001). Introduction to Linear Regression Analysis (3rd ed.). Wiley-Interscience.

Prati, G., & Mancini, A. D. (2021). The psychological impact of COVID-19 pandemic lockdowns: A review and meta-analysis of longitudinal studies and natural experiments. Cambridge University Press.

Rahman, M. M., & Davis, D. N. (2013). Addressing the Class Imbalance Problem in Medical Datasets. International Journal of Machine Learning and Computing.

Ratner, B. (2009). The correlation coefficient. Journal of Targeting, Measurement and Analysis for Marketing, 139–142.

Saarela, M., & Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. SN Applied Science, 3.

Singh, K., Upadhyaya, Dr. S., & Singh, S. N. (2012). Outlier Detection: Applications And Techniques. International Journal of Computer Science Issues, 307–323.

Varma K., V. (2020). Embedded methods for feature selection in neural networks. arXiv.

Weston, J., Poggio, T. A., & Mukherjee, S. (2000). Feature Selection for SVM. Neural Information Processing Systems Foundation, 55.

8 Appendix

8.1 Performance Measures

Table 8: SVM w/ SMOTE Results

	SMOTE							
	Min-Max Scalar				Standard Scalar			
	F2-Beta		Recall		F2-Beta		Recall	
	Pred_neg	Pred_pos	Pred_neg	Pred_pos	Pred_neg	Pred_pos	Pred_neg	Pred_pos
No Stroke	464	733	2	1195	483	714	9	1188
Stroke	0	80	0	80	0	80	1	79

Table 9: SVM w/ ADASYN Results

	ADASYN							
	Min-Max Scalar				Standard Scalar			
	F2-Beta		Recall		F2-Beta		Recall	
	Pred_neg	Pred_pos	Pred_neg	Pred_pos	Pred_neg	Pred_pos	Pred_neg	Pred_pos
No Stroke	158	1039	0	1197	0	1197	123	1074
Stroke	1	79	0	80	0	80	0	80

Table 10: RF w/ SMOTE Results

	SMOTE							
	Min-Max Scalar				Standard Scalar			
	F2-Beta		Recall		F2-Beta		Recall	
	Pred_neg	Pred_pos	Pred_neg	Pred_pos	Pred_neg	Pred_pos	Pred_neg	Pred_pos
No Stroke	776	421	0	1197	776	421	0	1197
Stroke	13	67	0	80	13	67	0	80

Table 11: RF w/ ADASYN Results

	ADASYN							
	Min-Max Scalar				Standard Scalar			
	F2-Beta		Recall		F2-Beta		Recall	
	Pred_neg	Pred_pos	Pred_neg	Pred_pos	Pred_neg	Pred_pos	Pred_neg	Pred_pos
No Stroke	776	421	0	1197	776	421	0	1197
Stroke	13	67	0	80	13	67	0	80

Table 12: MLP w/ SMOTE Results

	SMOTE							
	Min-Max Scalar				Standard Scalar			
	F2-Beta		Recall		F2-Beta		Recall	
	Pred_neg	Pred_pos	Pred_neg	Pred_pos	Pred_neg	Pred_pos	Pred_neg	Pred_pos
No Stroke	903	294	1	1196	964	233	1	1196
Stroke	29	51	0	80	38	42	0	80

Table 13: MLP w/ ADASYN Results

	ADASYN							
	Min-Max Scalar				Standard Scalar			
	F2-Beta		Recall		F2-Beta		Recall	
	Pred_neg	Pred_pos	Pred_neg	Pred_pos	Pred_neg	Pred_pos	Pred_neg	Pred_pos
No Stroke	625	572	1	1196	986	211	1	1196
Stroke	5	75	0	80	36	44	0	80

8.2 Outliers

Figure 3: Before removing outliers

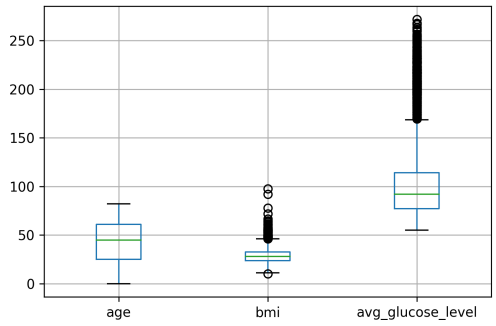
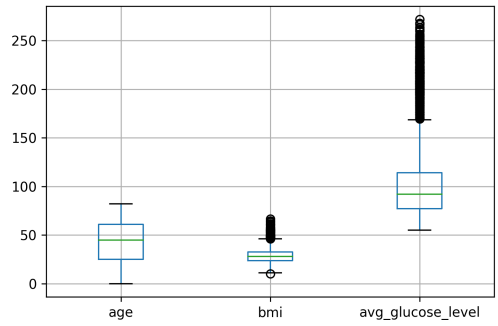


Figure 4: After removing outliers



8.3 Feature Importance

Figure 5: RF w/ SMOTE

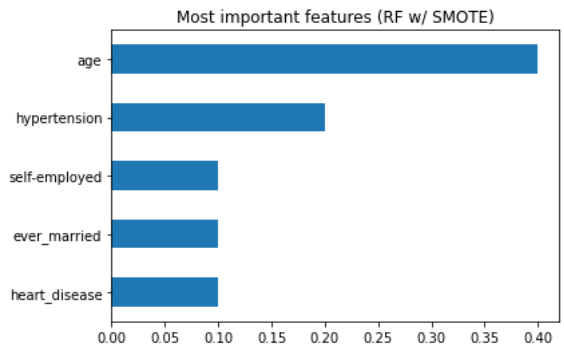


Figure 6: RF w/ ADASYN

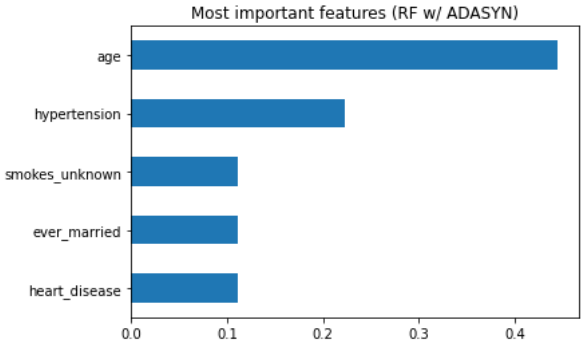


Figure 7: MLP w/ SMOTE

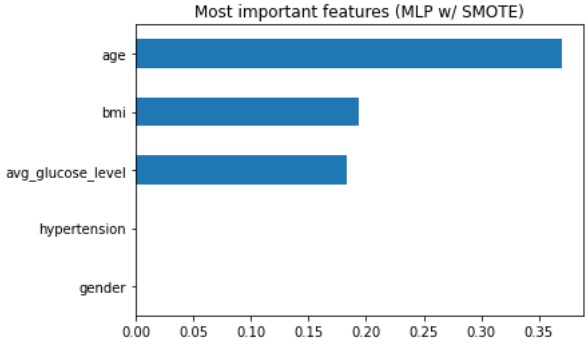


Figure 8: MLP w/ ADASYN

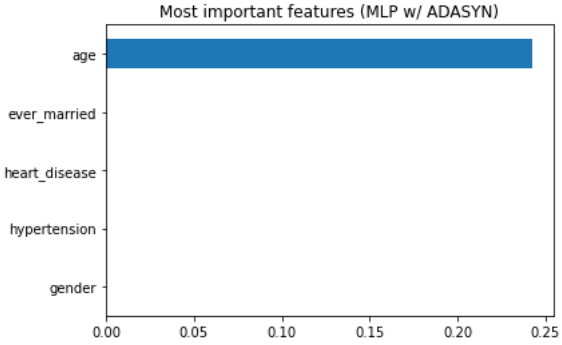


Figure 9: SVM w/ SMOTE

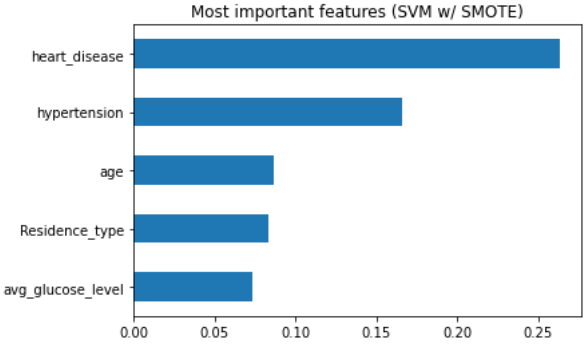


Figure 10: SVM w/ ADASYN