Stroke Analysis and Prediction Using PySpark, Support Vector Machine and Random Forest Regression

Aid Semić [a,1,*] , Sulejman Karamehić [a,2]

[a] International Burch University, Sarajevo, Bosnia and Herzegovina
[1] aid.semic@stu.ibu.edu.ba; [2] sulejman.karamehic@stu.ibu.edu.ba
* corresponding author

ARTICLE INFO

Article history
Received March 15, 2022
Revised August 7, 2022
Accepted October 17, 2022

Keywords
stroke
analysis
pyspark
suport vector
random forest
heart
blood vessels

ABSTRACT

Stroke is a medical condition in which the blood vessels in the brain rupture, causing brain damage. Symptoms may appear if the brain's flow of blood and other nutrients is disrupted. Stroke is the leading cause of death and disability worldwide, according to the World Health Organization (WHO). Early awareness of the numerous stroke warning symptoms can assist to lessen the severity of the stroke. To forecast the likelihood of a stroke happening in the brain, many machine learning (ML) models have been developed. This research uses a range of physiological parameters and machine learning algorithms, such as Support Vector Machine with extensive Exploratory Data Analysis, Random Forest Regression and PySpark. By using this methodologies and algorithms we got very high accuracy score results which are described down below.

1. Introduction

A myocardial infarction (commonly called a heart attack) is an extremely dangerous condition caused by a lack of blood flow to your heart muscle. The lack of blood flow can occur because of many different factors but is usually related to a blockage in one or more of your heart's arteries. Without blood flow, the affected heart muscle will begin to die. If blood flow isn't restored quickly, a heart attack can cause permanent heart damage and death.

How common are heart attacks? New heart attacks happen to about 635,000 people in the U.S. each year. About 300,000 people a year have a second heart attack. About one in seven deaths in the U.S. is due to coronary heart disease, which includes heart attacks. What happens during a heart attack? When a heart attack happens, blood flow to a part of your heart stops or is far below normal, which causes that part of your heart muscle to die. When a part of your heart can't pump because it's dying from lack of blood flow, it can disrupt the pumping sequence for the entire heart. That reduces or even stops blood flow to the rest of your body, which can be deadly if it isn't corrected quickly. [1]

What causes a heart attack? The vast majority of heart attacks occur because of a blockage in one of the blood vessels that supply your heart. This most often happens because of plaque, a sticky substance that can build up on the insides of your arteries (similar to how pouring grease down your kitchen sink can clog your home plumbing). That buildup is called atherosclerosis. Sometimes, plaque deposits inside the coronary (heart) arteries can break open or rupture, and a blood clot can get stuck where the rupture happened. If the clot blocks the artery, this can deprive the heart muscle of blood and cause a heart attack. Heart attacks are possible without a blockage, but this is rare and only accounts for about 5 percent of all heart attacks [2].

2. Related work

Heart diseases are the leading cause of mortality in the United States, accounting for 4 percent to 10 per- cent of all fatalities in those under the age of 45. Poor circulation or pump failure may cause heart failure in newborns, infants, toddlers, and adolescents. Academics have long been fascinated by the notion of using machine learning and data analysis to identify heart illnesses. Implementation of a Heart Disease Risk Prediction Model Using Machine Learning, Karthick, Thomas [3].

Karthick et al. [3] enhanced the study technique for predicting the probability of getting Cardio Vascular Disease (CVD) in individuals under the age of fifty (50). This technique may aid in the early detection of cardiac issues or attacks in individuals over 50, perhaps lowering or avoiding mortality. Based on age, gender, blood pressure, cholesterol, and pulse rate, Nour et al [4] utilized data mining to estimate a person's risk of heart disease. Data mining techniques such as Nave Bayes, K-Nearest Neighbors, Decision Tree Algorithm, and Neural Networks were used to do this. The disease was classified using the decision tree technique, and the probability was forecasted using the Gaussian algorithm.

A predictive analytics approach for stroke prediction using machine learning and neural networks. Soumyabrata Dev, Hewei Wang, Chidozie Shamrock Nwosu [5].

Because of the harmful effects of stroke on society, there has been a concentrated push to enhance stroke management and diagnosis. Caregivers provide chances for better patient management by methodically mining and storing patients' medical records as technology and medical diagnosis become more synergistic. As a result, it's critical to investigate the interdependence of these risk factors in patients' medical records and to comprehend their respective contributions to stroke prediction. The numerous components in electronic health records are methodically analyzed in this research for effective stroke prediction. We find the most relevant factors for stroke prediction using several statistical techniques and principal component analysis. Age, heart disease, average glucose level, and hypertension are the most critical indicators in recognizing stroke in patients, according to our findings. Furthermore, when compared to using all available input features and other benchmarking methodologies, a perceptron neural network using these four parameters delivers thehighest accuracy rate and lowest miss rate. We describe our findings using a balanced dataset builtusing sub-sampling approaches because the sample is significantly skewed in terms of the occurrence of stroke.

An integrated machine learning approach to stroke prediction. Aditya Khosla, Yu Cao, Junling Hu [6]. Stroke is the third largest cause of death in the United States and the leading cause of serious long-term disability. Stroke prediction accuracy is critical for early intervention and therapy. On the Cardiovascular Health Study (CHS) dataset, we compare the Cox proportional hazards model with a machine learning approach for stroke prediction. We look at the difficulties of data imputation, feature selection, and prediction in medical datasets in particular. We offer a unique automatic feature selection technique based on our suggested heuristic: conservative mean, which selects robust features. In comparison to the Cox proportional hazards model and the L1 regularized Cox feature selection technique, our suggested feature selection algorithm achieves a larger area under the ROC curve (AUC) when combined with Support Vector Machines (SVMs). We also describe a margin-based censored regression algorithm that combines the idea of margin-based classifiers with censored regression to produce a higher concordance index than the Cox model. In terms of AUC and concordance index, our approach exceeds the present state-of-the-art. In addition, our research has uncovered possible risk factors that have previously been overlooked by established methods. Our technique can be used to predict the clinical outcome of various diseases when incomplete data is widespread and risk factors are unknown.

3. Material and Methods

3.1 Dataset

The datasets that we have used in this work is Healthcare Stroke Dataset from kaggle. The author of this dataset is Rashik Rahman [7].The dataset consists of patients records from an undisclosed location with randomly selected patients which had or did not have a hart attack. It contains 5110 patients records with their medical information.

Table 1. Dataset attributes

Attribute	Description
Gender	Gender of the patient
Age	Age of the patient
Hypertension	Information does the patient have hypertension
Heart disease	Information about patient having heart disease
Ever married	Marital status information
Work type	Type of work the patient is obligated to
Residence type	Information where does the patient live
Average glucose level	Average glucose level
Bmi	Body mass index
Smoking status	Information is the patient a smoker or not
Stroke	Information about the patient stroke status

Each row in this dataset represent a person/patient with his medical information, table above is used to have proper foundings to better understand the data we are working with. The main goal of this work is to analize a large amount of data, for which purpose we have decided to use PySpark and following the analysis to create a prediction on a unknown patient.

3.2 PySpark

PySpark is the Python API for Apache Spark, an open source distributed computing platform and library for large-scale data processing in real time. PySpark is a good language to learn if you're already familiar with Python and libraries like Pandas. It'll help you construct more scalable analytics and pipelines. Apache Spark is a computational engine that works with large volumes of data in parallel and batch systems to process them. Spark is built in Scala, and PySpark was created to help Spark and Python work together. PySpark, in addition to providing a Spark API, uses the Py4j package to let you interact with Resilient Distributed Datasets (RDDs). The Spark dataframe is the most important data type in PySpark. This object functions similarly to dataframes in R and Pandas, and can be thought of as a table dispersed throughout a cluster. If you wish to use PySpark for distributed computation, you'll need to work with Spark dataframes rather than conventional Python data types. [8]

3.3 Support Vector Machine

The support vector machine algorithm's goal is to find a hyperplane in an N-dimensional space (N — the number of features) that distinguishes between data points. There are numerous hyperplanes from which to choose to split the two groups of data points. Our goal is to discover a plane with the greatest margin, or the greatest distance between data points from both classes. Maximizing the margin distance gives some reinforcement, making it easier to classify future data points. Hyperplanes are decision boundaries that aid in data classification. Different classes can be assigned to data points on either side of the hyperplane. The hyperplane's dimension is also determined by the number of features. The hyperplane becomes a two-dimensional plane when the number of input features reaches three. When the number of features exceeds three, it becomes impossible to imagine. Support vectors are data points that are closer to the hyperplane and have an influence on the hyperplane's position and orientation. We maximize the classifier's margin by using these support vectors. The hyperplane's position will be altered if the support vectors are deleted. These are the points that will assist us in constructing our SVM. [9]

3.4 Random Forest Regression

Random Forest Regression is a supervised learning approach for regression that use the ensemble learning method. The ensemble learning method combines predictions from several machine learning algorithms to produce a more accurate forecast than a single model. During training, a Random Forest constructs many decision trees and outputs the mean of the classes as the prediction of all the trees. Let's go over the steps to acquire a better knowledge of the Random Forest algorithm:

1. Pick at random k data points from the training set.
2. Build a decision tree associated to these k data points.
3. Choose the number N of trees you want to build and repeat steps 1 and 2.
4. For a new data point, make each one of your N-tree trees predict the value of y for the data point in question and assign the new data point to the average across all of the predicted y values.

Random Forest Regression is a powerful and precise model. It usually works well on a wide range of issues, including those with non-linear relationships. However, there are some drawbacks: there is no interpretability, overfitting is a possibility, and we must choose the amount of trees to include in the model. [10]

3.5 Proposed Method

For the experimental part of this work we have done the following. Since we are working with waste amounts of data we have decided to procces and format this data using PySpark which we have mentioned earlier, our data was in csv format and we decided by using PySpark to create the dataframe format of our data which will be used to mke predictions.

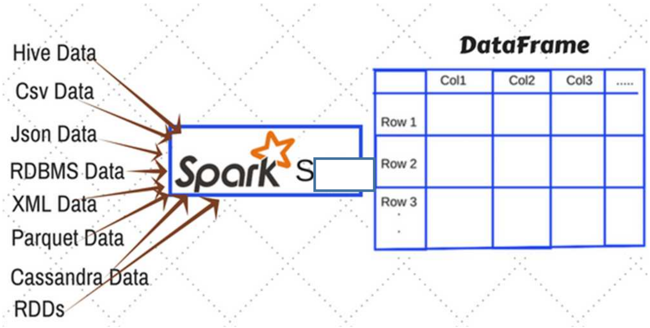


Figure 1. Representation of PySpark data processing

After we have created our data frame from our waste amount of data, we will be analzing this data and preforming various data preprocessing method. Following the analysis and preporcesing we are going to create two models which will be trained to predict from patient data if they have a liklyhood of having a heart stroke. Firstly we created a SVM model which we fited with testing data, and preformed prediction of reuslts on unknown data. On the image below SVM example graph is presented the closer the cluster to central vertic the better the result is for this model.

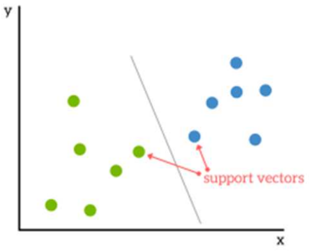


Figure 2. Support vector example

Following SVM model, we decided to use one more complex model which is Random Forest Regression, we first as previoulsy fitted the model with training data, and following that step we created the final output. Image below represents how a RFR creates the prediction from each nodes and generate a final output.

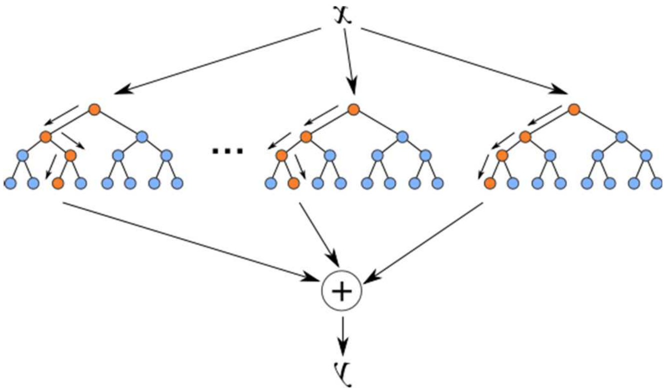
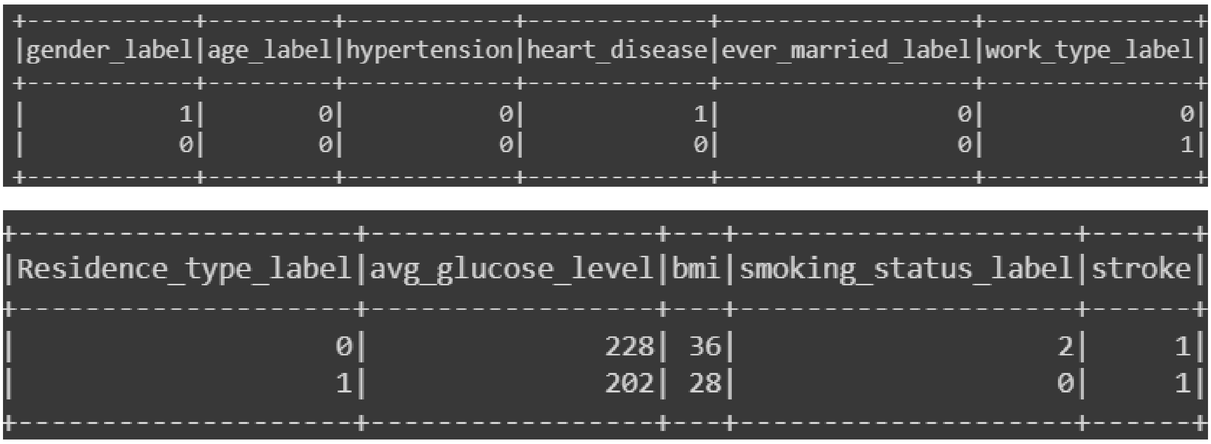


Figure 3. Random forest regression example

4. Results and Discussion

First step to get result was to perform PySpark on dataset and it was done successfully, we got the results and we were ready for algorithm implementation.



```
+------------+---------+------------+-------------+------------------+---------------+
|gender_label|age_label|hypertension|heart_disease|ever_married_label|work_type_label|
+------------+---------+------------+-------------+------------------+---------------+
|           1|        0|           0|            1|                 0|              0|
|           0|        0|           0|            0|                 0|              1|
+------------+---------+------------+-------------+------------------+---------------+
```

```
+--------------------+-----------------+---+--------------------+------+
|Residence_type_label|avg_glucose_level|bmi|smoking_status_label|stroke|
+--------------------+-----------------+---+--------------------+------+
|                   0|              228| 36|                   2|     1|
|                   1|              202| 28|                   0|     1|
+--------------------+-----------------+---+--------------------+------+
```

Figure 4. Perform PySpark on dataset

After performing Support Vectom Machine on dataser we got following results.

Table 2. Results of SVM

	Precision	Recall	F1 score	Support
0	0.95	0.81	0.88	1207
1	0.12	0.45	0.20	71
Accuracy			0.79	1278
Macro avg	0.54	0.63	0.54	1278
Weighted avg	0.92	0.79	0.84	1278

Accuracy score is equal to ***0.7942097026604069*** and we can say that this is very good result. This means that model fits aproximatley 79.4%. F1 score is equal to ***0.19571865443425074.*** This means that balance between the precision and the recall is 19.5%.

Now let's see the plot that which is used to explain prediction.

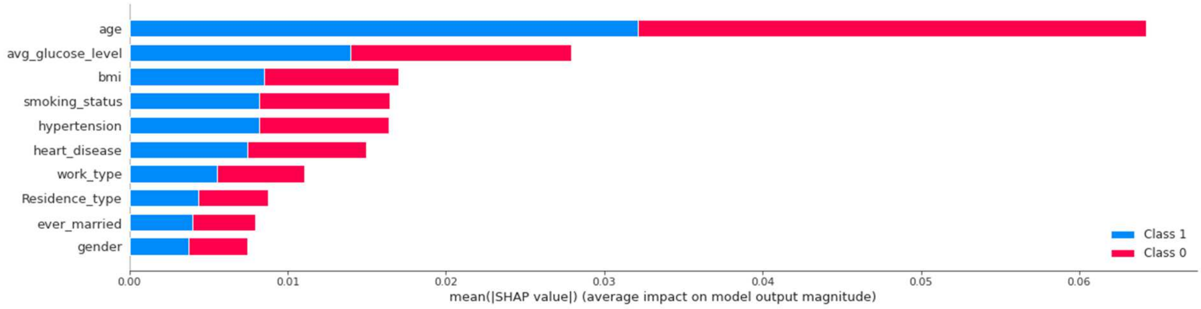


Figure 5. Results of SVM prediction

The graph above (Figure 5) shows the features used in predicting the stroke probability. Age appears to be on top which has high corelation with the Stroke probability, followed by Average glucose level and BMI.

While implementing Random Forest Regression interesting thing is that there were no correlation between features.

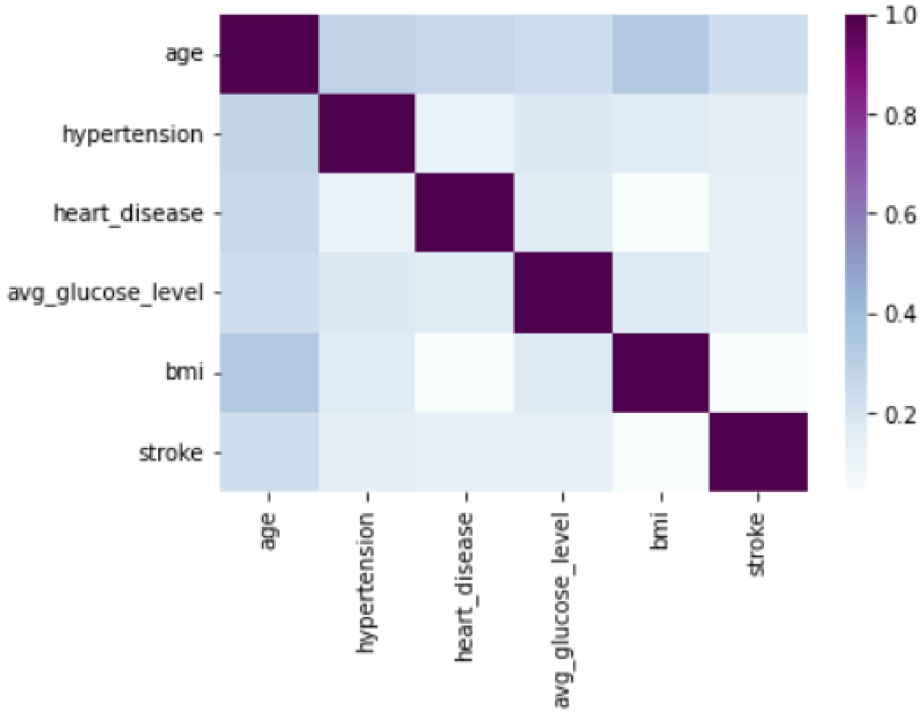


Figure 6. Correlation plot

After encoding and fixing all problems we got prefect results. Accuracy score was 100%.

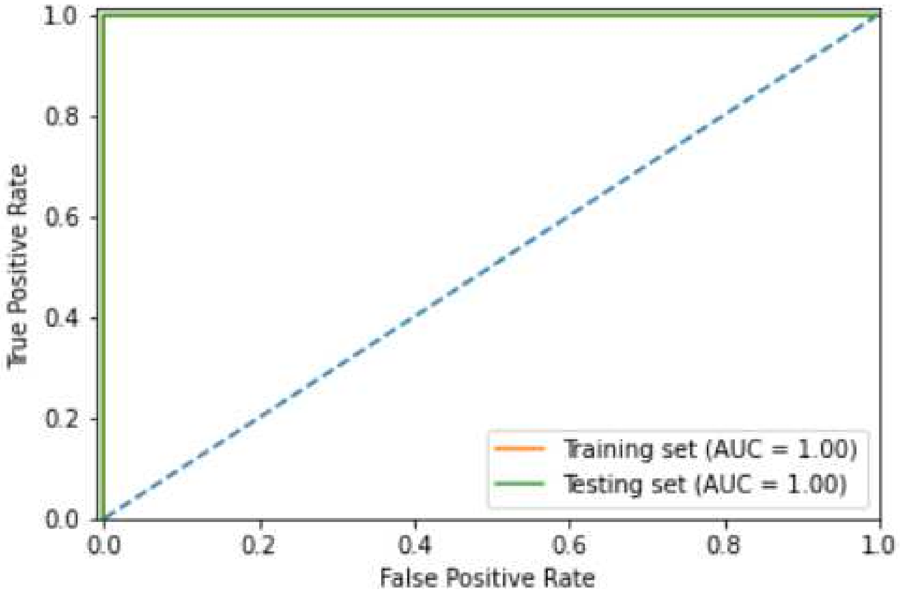


Figure 7. Random Forest Regression result

To show more of this accuracy let's see a confusion matrix of this algorithm.

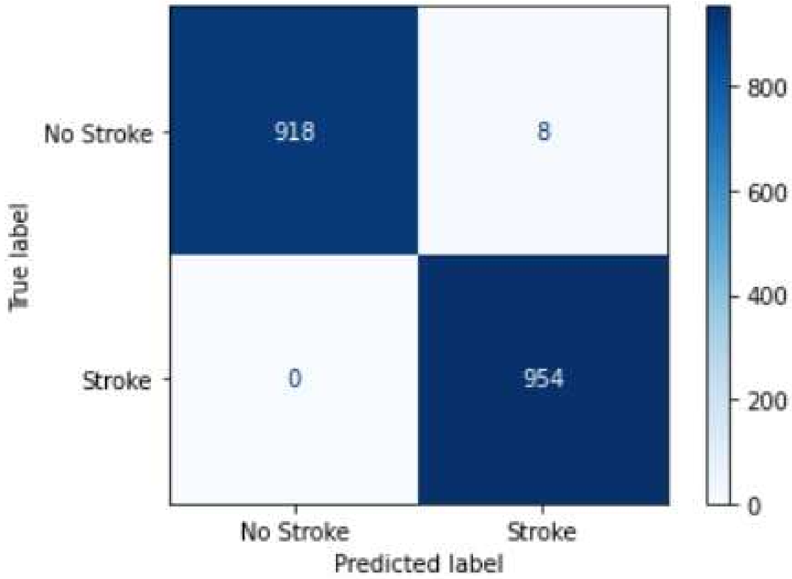


Figure 8. Random Forest Regression confusion matrix

In this paper, two machine learning techniques have been used for prediction of stroke at patients, besides that pyspark was used for data analysis and data exploration processes, but the main question is how good the predictions were since this is a sensitive topic, high precision is mandatary in these cases and this paper has proved that Random Forest Regression has performed with quite unreal accuracy with 100% this may be to a few factors type of dataset or the sheer size of the data, but nonetheless it has proven that it is the best model for this type of predictions as in this paper another machine learning technique was used which is Support Vector Machine and it has produced a quite big difference in the result data as it has produced around 79% accuracy which is still a very good score but in these cases as we have research anything belowe 90% accuracy is not good enough. To further more check this difference between these two machine learning techniques a research has been done in related works with the same two techniques to see was there the same difference as in this paper. Harshitha and colleagues [11] have done extensive research on many machine learning techniques and in their work RFR has produced 91.2% accuracy while SVM has produced 81.4% accuracy, this confirms the dominance of RFR against SVM as it was the case in this paper, besides them Rodriguez [12] has done research on the same two techniques but in his work the results for RFR very low at 72% accuracy while SVM outperformed RFR with 78.6% accuracy, this was the only work where SVM has produced a better results, but there a lot of reasons for this but the main one is the data as the data is the most important part of the whole process no matter how good of a training we do on the classifier if the data is bad we will still have bad results. Gangavarapu and Kumari [13] have done extensive research on the following techniques, and have produced the following results RFR 95.5% and SVM 92.4% accuracy, from their results we can see that RFR has again shown better results than SVM, but not a big difference possibly of some better optimization of SVM, bur regardless our results are in line with these ones as RFR has been proven to work better in these cases than SVM, but as mentioned earlier we can never underestimate the importance of the data we are working with.

5. Conclusion

This paper describes how PySpark may be used to set up and carry out data analysis on a selected dataset. On this dataset, PySpark was demonstrated along with the implementation of algorithms to obtain accuracy scores. Implementation process was fully successful. For better results it is necessary to conduct more research to determine whether the PySpark sorting module can be enhanced to remove the two-column restriction and make it theoretically customisable for any input file. We got almost perfect result by Random Forest Regression. Furthermore, future research may use other training, testing, and accuracy algorithms to get better picture of this topic.

References

[1] J. Lawton, “Ground-breaking discoveries in cardiovascular diseases”.

[2] P. P. Weissberg, “The future of cardiovascular research”.

[3] Khan SU, Kalra A, Yedlapati SH, Dani SS, Shapiro MD, Nasir K, et al. Stroke-Related Mortality in the United States–Mexico Border Area of the United States, 1999 to 2018. Journal of the American Heart Association [Internet]. 2021 Jul 6;10(13). Available from: http://dx.doi.org/10.1161/jaha.120.019993.

[4] Alsmadi T, Alqudah N, Najadat H. Prediction of Covid-19 patients states using Data mining techniques. 2021 International Conference on Information Technology (ICIT) [Internet]. 2021 Jul 14; Available from: http://dx.doi.org/10.1109/icit52682.2021.9491716.

[5] Dev S, Wang H, Nwosu CS, Jain N, Veeravalli B, John D. A predictive analytics approach for stroke prediction using machine learning and neural networks. Healthcare Analytics [Internet]. 2022 Nov;2:100032. Available from: http://dx.doi.org/10.1016/j.health.2022.100032

[6] Khosla A, Cao Y, Lin CC-Y, Chiu H-K, Hu J, Lee H. An integrated machine learning approach to stroke prediction. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’10 [Internet]. 2010; Available from: http://dx.doi.org/10.1145/1835804.1835830.

[7] R. Rahman, “Heart Stroke Dataset”.

[8] “Spark Apache,” [Na mreži]. Available: https://spark.apache.org/docs/latest/api/python/.

[9] “SciKitLearn,” [Na mreži]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[10] “Towards Data Science,” [Na mreži]. Available: https://towardsdatascience.com/random-forest-regression-5f605132d19d.

[11] H. P. G. G. V. P. P. K. B. Harshitha K V, “Stroke Prediction Using Machine Learning Algorithms,” International Journal of Innovative Research in Engineering & Management (IJIREM), tom 8, br. 4, 2021.

[12] J. A. T. Rodríguez, “Stroke prediction through Data Science and Machine Learning Algorithms”.

[13] G. L. A. K. Gangavarapu Sailasya, “Analyzing the Performance of Stroke Prediction using ML Classification Algorithms,” (IJACSA) International Journal of Advanced Computer Science and Applications, tom 12, br. 6, 2021.