Exploratory Data Analysis and Machine Learning Algorithms to Classifying Stroke Disease

Prismahardi Aji Riyantoko[1]
Department of Data Science
UPN "Veteran" Jawa Timur
Surabaya, Indonesia
prismahardi.aji.ds@upnjatim.ac.id

Tresna Maulana Fahrudin[2]
Department of Data Science
UPN "Veteran" Jawa Timur
Surabaya, Indonesia
tresna.maulana.ds@upnjatim.ac.id

Kartika Maulida Hindrayani[3]
Department of Information System
UPN "Veteran" Jawa Timur
Surabaya, Indonesia
kartika.ds@upnjatim.ac.id

Mohammad Idhom[4]
Department of Data Science
UPN "Veteran" Jawa Timur
Surabaya, Indonesia
idhom@upnjatim.ac.id

***Abstract*— This paper presents data stroke disease that combine exploratory data analysis and machine learning algorithms. Using exploratory data analysis we can found the patterns, anomaly, give assumptions using statistical and graphical method. Otherwise, machine learning algorithm can classify the dataset using model, and we can compare many model. EDA have showed the result if the age of patient was attacked stroke disease between 25 into 62 years old. Machine learning algorithm have showed the highest are Logistic Regression and Stochastic Gradient Descent around 94,61%. Overall, the model of machine learning can provide the best performed and accuracy.**

***Keywords*— eda, stroke, machine learning, classification**

I. INTRODUCTION

brain is interrupted or reduced, preventing brain tissue from getting oxygen and nutrients, which cause brain cells begin to die in minutes. A stroke is a medical emergency, and prompt treatment is crucial. Early action can reduce brain damage and other complications. Stroke was classified as a disease of the blood vessels by International Classification of Disease 11 (ICD-11) was released in 2018 [1]. Stroke to be most popular disease in the world because stroke is in the second leading cause of death globally. Almost 13.7 million people affected and kills around 5.5 million by stroke disease. Stroke are classified as ischemic which have two types, there are thrombotic and embolic [2]. The thrombotic stroke has symptom the blood clot forms in one of the arteries if supplies blood to human brain. The embolic have symptom a blood clot forms away from the patient bloodstream to lodge in narrower brain arteries.

This paper reviewed and analyzed the current studies on classification of stroke using exploratory data analysis and machine learning approach. Exploratory data analysis refers to the critical process on data to discover patterns, anomaly, give a hypothesis to check assumptions using statistical method and graphical representations. Currently, we will focus on stroke disease which have category like gender, age, hypertension, heart disease, ever married, work type, residence type, average of glucose, body mass index (BMI), and smoking status. The category of stroke disease based on healthcare dataset we gain from Kaggle.

In several countries has increasing a large number of people lose their life due to stroke disease [3]. The stroke disease has high risk factor based on the type. The machine learning algorithm can improve patient healthy through to detection and classification. Furthermore, in this case has developed a classification model for stroke using Logistic Regression, Naïve Bayes, Random Forest, Extreme Gradient Boost, Gradient Boost, K-Nearest Neighbor (k-NN), Decision Tree, Support Vector Machine and Stochastic Gradient Descent. The classification model is based on a dataset of 5110 cases collected from healthcare dataset stroke in Kaggle.

II. RELATED WORKS

In this section we will describes the previous research about stroke disease which analyzed use exploratory data analysis and machine learning algorithms. In the recent years, exploratory data analysis and machine learning algorithms has different works in published, therefore we will discuss about that here. Shoily et al, used machine learning algorithms for detecting of stroke disease, and they used Naïve Bayes, J48, k-NN, and Random Forest to trained models [4]. They have collected the data from various sources with total 1058 individual patient's data information with two type gender for male are 412 data and for female are 646 data. Based on the classification model, J48, k-NN, and Random Forest have the highest accuracy with 98,8% and Naïve Bayes have accuracy 85,6% for classifying

the stroke disease. The novelty in this paper is contributing their networks for collecting and preparing dataset using WEKA.

Adam, Yousif and Bashir, using Machine Learning Algorithms for classifying of Ischemic Stroke, and they using performance of decision tree classification tree is better than k-NN algorithm. The results help the medical officer for classifying of ischemic stroke.

Cheon, Kim, and Lim, using deep learning to predicted stroke patient mortality for Korean population. They have collected and used around 15.099 data of patient of stroke. The dataset was extracted using Principal Component Analysis (PCA) to gaining relevant background features from medical records. They used deep learning approach and compared with five other machine learning methods. Based on the results, value of Area Under Curve (AUC) was 83.48%, hence the methods can be used by the doctors to predict and gain the hypothesis to detecting the stroke in human body.

In the other research, Indrakumari and their colleague using exploratory data analysis to predict heart disease, and they have used big data to gain the information, pattern, and knowledge for decision making [6]. The Exploratory Data Analysis (EDA) can be used to detects some mistakes, find appropriate data, assumptions, and determines the correlations. They have used K-means algorithm to predicting and analyses heart disease with 209 data records. Based on the data, they consist 8 category data like age, chest pain type, blood pressure, blood glucose level, EGG on rest, heart rate, and four types of chest pain. In the final results, this research show that the prediction was gave an accurate value.

III. METHODOLOGY

In this section, we will describe the methodology to pre-processing and processing the dataset using exploratory data analysis and machine learning algorithm for classifiers stroke disease.

A. Data Sources

The data sources of stroke disease have collected from Kaggle. Our dataset contains many categories such are gender, age, hypertensions, heart disease, ever married, work type, residence type, average glucose level, body mass index (BMI), and smoking status. The data sources contain 5110 data of stroke disease patient. In the other hand the data of BMI only 4909 data, so this anomaly condition from the data sources.

TABLE I. LIST OF CATEGORIES OF THE DATA SOURCES

SI	List of Categories	
	Attributes	***Description***
1	Gender	Gender of Patient (Male and Female)
2	Age	Age of Patient
3	Hypertensions	A situation for high blood pressure of patient
4	Ever Married	A condition of patient has marital status
5	Work Type	A work condition in the office of patient like private employed or self-employed
6	Residence Type	A place of patient which related with their live or work area
7	Average Glucose Level	A condition of patient which have normal or abnormal blood sugar ranges
8	Body Mass Index (BMI)	A measure of body fat based on height and weight of patient
9	Smoking Status	A condition of smoker or no from patient

decision tree and K-Nearest Neighbor (k-NN) [2]. They collected the data based on a dataset of 400 cases from different Sundanese Hospital. Based on the experiment, the

B. Exploratory Data Analysis

The exploratory data analysis we will describes about the statistical and graphical or non-graphical using categories in the dataset. Furthermore, EDA was used to analyze the data with advanced technique to gain information such us expose hidden structure, enhances the insight, identifies anomalies, and builds parsimonious model to test the underlaying assumptions [6]. In addition, we will provide to calculate the statistical formulation such as mean, standard deviation, and quartile.

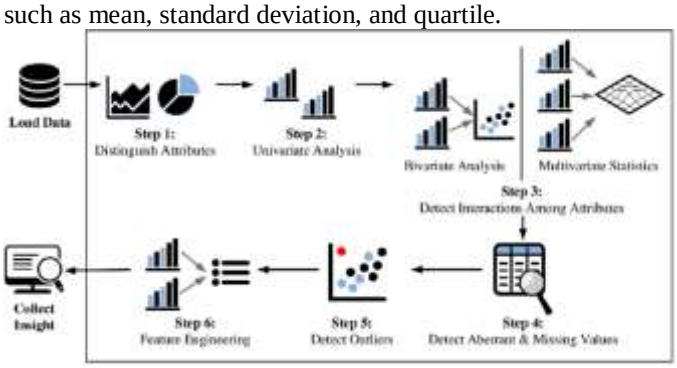

Fig 1. Diagram Exploratory Data Analysis

C. Machine Learning Algorithms

In this works, machine learning algorithm will be use to classifying the stroke disease according to dataset. The following will provide an describe about classification algorithm.

1) Logistic Regression,

Logistic regression is a analytical statistics method for describe the correlations between independet variable which have two type categories or more [7]. The model form Predicted Probabilities of Logistic Regression which describe using natural logarithm as follow

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x \qquad (1)$$

The model of logistic regression will be uses to analyze correlation between one variable of reponses and many predictor variable, with the response varieble in the form dichotomous qualitative data that have value 1 for describe existence of characteristics and value 0 for describe non-existence of characteristics [8].

2) Naïve Bayes,

Naïve bayes was used for classification data based on Bayes Theorem. The Bayes theorem have function to finds

conditional probability in each event. The Bayes form can be formulate as follows

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2)$$

Bayesian classifier used for handling the dataset containing many attributes. The advantage of bayesian, the classification model is simple, and have rapidity of use.

3) Random Forest,

Random forest method is development algorithm from CART method there are adjust bootstraping aggregating method and random feature selection. In this classfier method, there are many decision tree hence build a forest, subsequently the tree will be analyze.

4) Gradient Boost and Extreme Gradient Boost,

Xtreme Gradient Boosting (XGBoost) is a great method combine boosting with gradient boosting. This method the first time introduced by Friedman, in his research used the relationship between boosting and optimization to create a Gradient Boosting Machine (GBM). Model built using the boosting method, namely by making a modelnew to predict the error from the previous model.

XGBoost is a version of the Gradient Boosting Method (GBM) more efficient and scalable because it is able to complete various functions such as regression, classification, and ranking. XGBoost was first introduced to the Higgs Boson Competition, where in this competition the XGBoost method becomes a method most used by most teams. Apart from these competitions, XGBoost method is also a method that is widely used in competition machine learning organized by Kaggle in 2015. XGBoost is a tree ensembles algorithm consisting of several classification and regression trees (CART). The XGBoost algorithm performs 10 times more optimization fast compared to implementations of other GBMs [9].

5) K-Nearest Neighbor (k-NN),

K-Nearest Neighbor Algorithm is a method of classification grouping new data by distance new data to several data / neighbors [11]. K-NN have a framework start from to training data, labeling data, and testing data. To determine the distance between x and y can use Euclidean equation as follows

$$d(X_1, Y_1) = \sum_l \left| \frac{n_{1i}}{n_1} - \frac{n_{2i}}{n_2} \right| \quad (3)$$

6) Decision Tree,

Decision Tree is used for

study the classification and prediction of patterns from data and describe the relation of the variable attribute x and the target variable y in the form of a tree [12]. The Decission Tree is a structure like flowchart where each internal node (node which is neither leaf nor outermost node) represents testing of the attribute variable, each the branch is the result of testing that, while the outermost node is the leaf became the label [13].

7) Support Vector Machine

The Support Vector Machine (SVM) was developed by Boser, Guyon, and Vapnik was first presented in 1992 at the Annual Workshop on Computational Learning Theory. The basic concept of SVM is a combination of computational theories that had existed in previous years were like margins hyperplane [10]. The way SVM works is to find a dividing field called the best hyperplane that divides the data into 2 different classes. In its development, SVM can be expanded to a classification of more than two classes or multiple classes. Unlike ANN, SVM will determine the most optimum hyperplane where the hyperplane is said to be optimum if it is right in the middle of the two classes so it has the most distance far to the outer data in both classes.

8) Stochastic Gradient Descent

SGD method is an iterative optimization algorithm to find the minimum function point that can be lowered. The algorithm starts with a do the logging early in the process [14].. Error the logs were then corrected over time there is a loop of guesses using the rules gradient (derivative) of the desired function minimized. Minimum function inheritance is used specifically with formula as follow

$$\omega_i + 1 = \omega_i - \eta \nabla \omega_i L(\omega_i) \quad (4)$$

IV. THE RESULTS AND DISCUSSION

In this section we will provide results exploratory data analysis and machine learning algorithm for classification.

A. Exploratory Data Analysis

Based on the data sources, we have extracted the data and get 5110 rows and 11 columns.

TABLE II. LIST OF DATA SOURCES BASED ON COLUMNS AND ROWS

SI	List of Data		
	Attributes	***Unique Values***	***Missing Value***
1	Gender	3	0
2	Age	104	0
3	Hypertensions	2	0
4	Ever Married	2	0
5	Work Type	5	0
6	Residence Type	2	0
7	Average Glucose Level	3979	0
8	Body Mass Index (BMI)	418	201
9	Smoking Status	4	0

In the Table II, we have provided two type values, there are unique values and missing values based on dataset. We get anomaly data in the BMI column missing value, it showed that sustained missing value 201 data.

	count	mean	std	min	25%	50%	75%	max
age	5110.0	43.226614	22.612647	0.08	25.000	45.000	61.00	82.00
hypertension	5110.0	0.097456	0.296607	0.00	0.000	0.000	0.00	1.00
heart_disease	5110.0	0.054012	0.226063	0.00	0.000	0.000	0.00	1.00
avg_glucose_level	5110.0	106.147677	45.283560	55.12	77.245	91.885	114.09	271.74
bmi	4909.0	28.893237	7.854067	10.30	23.500	28.100	33.10	97.60
stroke	5110.0	0.048728	0.215320	0.00	0.000	0.000	0.00	1.00

Fig 2. Statistical Value in Mean, Standard Deviation, and Quartiles

In the figure 2, output for numerical data EDA shows some simple distribution statistics that include mean, standard deviation, and quartiles. Some of the points that we can drive from the output include if there no negative values for all numeric data. If EDA data be found negative value, it may mean that we have further investigate, and we may have to clean the data.

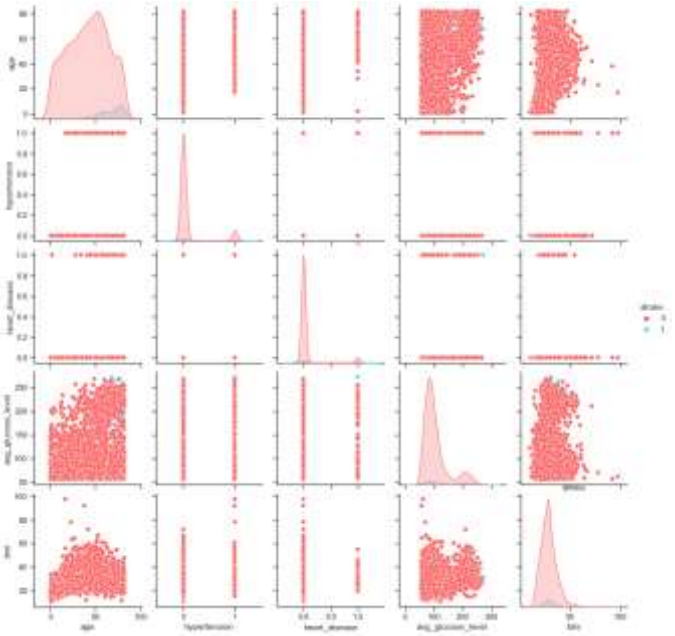

Fig 3. Joint Distribution Pairplot

The last output for numeric data EDA is a pairwise joint distribution plot. The graphical noticed that stroke disease is mostly distributed their age between approximately 25 into 62 years. In the other hand, people without hypertension have more risk to have a stroke, likewise people without any previous heart disease have more risk to have a stroke.

B. Machine Learning Algorithm for Classification

Of the various setting tested, we have nine model classification to testing using stroke disease dataset. In data processing from 5110 data, we use 3832 data for train, and 1278 data for test.

TABLE III. LIST OF CLASSIFICATION MODEL

SI	ML Algorithm	List of Machine Learning Algorithm for Classification			
		Confusion Matrix			
		TP	FP	FN	TN
1	Logistic Regression	1208	0	69	1
2	Naïve Bayes	195	1013	0	70
3	Random Forest	1208	0	70	0
4	Extrem Gradient Boost	1208	0	70	0
5	Gradient Boost	1206	2	70	0
6	K-Nearest Neighbor	1208	0	70	0
7	Decision Tree	1208	0	70	0
8	Support Vector Machine	1207	1	70	0
9	Stochastic Gradient Descent	1208	0	69	1

True Positive (TP) is predicted patient of stroke disease. True Negative (TN) is predicted patient have not stroke disease. False Positive (FP) is wrong predicted if the patient have not stroke disease. False Negative (FN) is wrong predicted if the patient have stroke disease. The Table III used for calculate accuracy, precision, recall, and F-1 score.

TABLE IV. LIST OF CLASSIFICATION MODEL

SI	List of Machine Learning Algorithm for Classification	
	ML Algorithm	Accuracy (%)
1	Logistic Regression	94,61
2	Naïve Bayes	20,74
3	Random Forest	94,53
4	Extrem Gradient Boost	94,53
5	Gradient Boost	94,36
6	K-Nearest Neighbor	94,53
7	Decision Tree	94,53
8	Support Vector Machine	94,45
9	Stochastic Gradient Descent	94,61

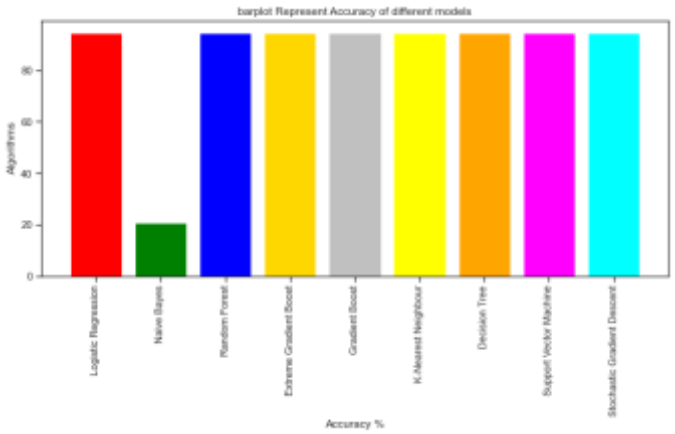

Fig 4. Accuracy of Machine Learning Algorithm

In the Table IV and Figure 4 have showed if all algorithms performed very similar but we will divide into two category, the highest accuracy with logistic regression and stochastic

gradient descent about 94,61%. In the other condition, the lowest accuracy with Naïve Bayes about 20,74%.

V. CONCLUSSION

In this paper, a sufficiently medium dataset of stroke attacked patients has been classified accurately based on exploratory data analysis and machine learning algorithm for classification. EDA can show the best result using statistical and graphical data analysis, we get anomaly data in BMI with missing 201 data. But in the other hand, the missing condition not significant to affecting the dataset. Overall, the EDA can show if the age patient of stroke disease between 25 until 62 years. Furthermore, people haven't hypertension and heart disease, potentially can attacked by stroke disease. The algorithm of machine learning for classification showed if the logistic regression and stochastic gradient descent give highest accuracy about 94,61%. Almost all model can reach more than 90%, but Naïve Bayes only has accuracy about 20,74%. For the future works, we need to gaining more information why the Naïve Bayes can't get the highest accuracy, because the Naïve Bayes have anomaly condition, maybe in the dataset can't support the model or vice versa.

REFERENCES

[1] Kuiakose, D and Xiao, Z (2020) "Pathophysiology and Treatment of Stroke: Present Status and Future Perspectives", International Journal of Molecular Sciences. MDPI, 21, 7609. Doi:10.3390/ijms21207609.

[2] Adam, SY., Yousif, A., and M.B. Bashir, (2016) "Classification of Ishemic Stroke Using Machine Learning Algorithms". International Journal of Computer Applications (0975-8887). Vol 149.

[3] L.T. Kohn, J. Corrigan, M.S. Donaldson, et al. To err is human: building a safer healt system. Vol 6. Naiional Academy Press Washington, DC. 2000.

[4] Shoily, T.I., et al. (2019). "Detection of Stroke Disease using Machine Learning Algorithms". 10th ICCCNT, July 6-8 2019 – IIT – Kanpur, India

[5] Cheon, S., Kim, J., and Lim, J. (2019). "The Use of Deep Learning to Predict Stroke Patient Mortality". Intrnational Journal of Environment Research Public Health (MDPI).

[6] Indrakumari, R. Poongodi, T., and Jena., R.J., (2020). "Heart Disease Prediction using Exploratory Data Analysis". International Conference on Smart Sustainable Intelligent Computing and Applications under ICTETM 2020. Procedia Computer Science 173. Page 130-139.

[7] Hosmer, D.W., and S. Lemeshow (2000). "Applied Logistic Regression". 2th Edition. John Wiley and Sons Inc, Canada.

[8] Sepang, F., H. Komalig, D. Hatidja (2012). Penerapan Regresi Logistik untuk Menentukan Faktor-faktor yang mempengaruhi Pemilihan Alat Kontrasepsi di Kecamatam Modayag Barat. Universitas Sam Ratulangi. Manado. Jurnal MIPA Unsrat Online 1(1): 1-5.

[9] Chen T., Guestrin C. (2016): XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August

[10] Vapnik, & N, V. (1999). The Nature of Statistical Learning Theory 2nd. New York Berlin Heidelberg: Springer

[11] Santoso, B. (2007). Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis (1 ed.). Yogyakarta: Graha Ilmu.

[12] Ye, N. (2014). Data Mining Theories, Algorithms, and Examples. 6000 Broken Sound Parkway NW: Taylor & Francis Group, LLC.

[13] Han, J., Kamber, M., & Pei, J. (2012). Data Mining Concepts and Techniques (3rd Editio). Waltham, USA: Morgan Kaufmann Publishers.

[14] A. S. Ritonga and E. S. Purwaningsih, "Penerapan Metode Support Vector Machine (SVM) Dalam Klasifikasi Kualitas Pengelasan Smaw (Shield Metal Arc Welding)," Ilm. Edutic, vol. 5, no. 1, pp. 17–25, 2018.