

Heart Stroke Prediction: A Comparative Analysis of RapidMiner and Colab using Neural Networks for Practical Application

John Carlo Astoveza
College of Computer Science and
Engineering
Jose Rizal University
Mandaluyong, Philippines
johncarlo.astoveza@my.jru.edu

Jasmin Clores
College of Computer Science and
Engineering
Jose Rizal University
Mandaluyong, Philippines
jasmin.clores@my.jru.edu

Evalyn Grace Estrera
College of Computer Science and
Engineering
Jose Rizal University
Mandaluyong, Philippines
evalyngrace.estrera@my.jru.edu

Mark Joshua Lodriga
College of Computer Science and
Engineering
Jose Rizal University
Mandaluyong City, Philippines
mark.lodriga@my.jru.edu

Brixter Luquing
College of Computer Science and
Engineering
Jose Rizal University
Mandaluyong City, Philippines
brixter.luquing@my.jru.edu

Mhargielyn Miñeque
College of Computer Science and
Engineering
Jose Rizal University
Mandaluyong City, Philippines
mhargielyn.mineque@my.jru.edu

Darwin Puzo
College of Computer Science and
Engineering
Jose Rizal University
Mandaluyong City, Philippines
darwin.puzo@my.jru.edu

Abstract— This study examines heart stroke prediction, conducting a thorough comparative analysis of the prevalent data science tools RapidMiner and Colab. Neural networks are utilized for practical implementation. The critical need for accurate and effective predictive models in healthcare drives the exploration. A review of current literature highlights the importance of predicting heart strokes, focusing on the risk of false positive and false negative predictions in medical diagnostics. Furthermore, the multiple studies emphasize the value of neural networks in handling intricate patterns within health data, which highlights their effectiveness in predictive modeling. The primary objective of this research is to identify the differences in experimental outcomes when implementing neural networks through Colab and RapidMiner. By combining insights from existing literature and conducting thorough experiments. This study aims to contribute valuable findings that inform the choice of data science tools for heart stroke prediction that optimize practical applications in real-world healthcare scenarios.

Keywords— Logistic Regression, Neural Networks, RapidMiner, Python Colab

I. INTRODUCTION

Logistic regression is a technique to predict the relationship between a binary response to a variable and one or more independent variables [1]. Logistic regression utilizes a logistic

function to convert the linear combination of explanatory factors into likelihood. The resulting change is essential for dealing with the different features of binary outcome that ensures the predicted probabilities belong to the range of 0 and 1. This method is a reliable statistical technique used to assess how a group of independent variables impact a binary result [2]. It is also particularly effective in modeling scenarios where the dependent variable represents a binary outcome, such as success or failure, presence or absence.

Logistic regression contains numerous limitations. The inclusion of small numbers and limited information might result in possible bias, which in turn can lead to incorrect judgments [3]. In this context, "bias" refers to the tendency of the logistic regression model to provide estimates that differ from the actual values due to the limitations caused by a small number of data variables. This constraint emphasizes the significance of thoroughly examining and assessing different modeling methods in situations.

The objective of this study is to use logistic regression to analyze how a group of explanatory variables affects a binary result, focusing on optimizing efficiency and effectiveness. The binary outcome can include a wide array of phenomena, encompassing from health problems to consumer behaviors.

Logistic regression offers a measurable method to determine each independent variable's distinct impact within a framework considering many variables.

II. LITERATURE REVIEW

Stroke Classification Model using Logistic Regression

The paper aims to determine what factors play a big role in categorizing different types of heart strokes. They used logistic regression to develop a model that can predict whether a patient is prone to experiencing a non-hemorrhagic stroke (NHS) or a hemorrhagic stroke (HS) [4]. The prediction is based on factors such as cholesterol levels, blood sugar levels, body temperature, duration of hospital stay, pulse rate, and gender. The paper dataset instances are 261 medical records from Dadi Hospital in Makassar. The dataset has cases of non-hemorrhagic stroke, with 161 cases and 100 cases of hemorrhagic stroke. The model that they have developed achieved an accuracy of 74.8% in classifying the heart stroke. The results indicated that cholesterol level and length of stay were the two significant predictors affecting the type of heart stroke. The study also performed simultaneous and partial significance tests to evaluate the overall and patient effects of the predictor variables on the heart stroke classification.

Logistic Regression Model based on heart disease and Its Potential Influencing Factors

The paper discusses how more people are getting heart disease, and its goal is to develop a model using logistic regression. This model will predict how likely people are to have heart disease by looking at five factors that might influence it, and this model may be used to help stop heart disease early on and treat it on time [5]. The dataset used contains 462 instances and has variables such as systolic blood pressure, family history, cholesterol levels, obesity measures, and age. The model's accuracy, with cross-validation methods, is reported to be over 73%. The paper concludes that the chosen factors have significant impacts on heart disease prediction and that the model can provide a preliminary prediction of a patient's risk of developing heart disease.

A Computational Model for Prediction of Heart Disease Based on Logistic Regression with Grid Search

The main goal of the research is to create a model for heart disease detection utilizing logistic regression and patient data already in existence. In comparison to previous research, the preprocessing and tuning methods are different [6]. The technique improves algorithm efficiency by adding new properties to certain features. Four classifiers are used in the implementation; the maximum accuracy is 93% for the logistic regression and sklearn functions. The study makes the claim that attribute significance might have an impact on mistakes. To further boost accuracy, future advancements may entail decreasing characteristics and using more effective algorithms.

Heart Disease Prediction Using Logistic Regression Algorithm

This study used the Heart Disease UCI dataset, which consisted of fourteen variables, to evaluate the efficacy of the logistic regression method in predicting cardiovascular disease [7]. The findings demonstrated an accuracy rate of 85% and a negligible error rate of 0.1406565, signifying the algorithm's efficacy in prediction. The research determined that gender, blood pressure level (trestbps), heart rate (thalach), and the number of afflicted arteries are major variables that impact the probability of developing heart disease. An elevation in these factors is linked to a decline in overall cardiovascular function and an escalation in the susceptibility to cardiovascular ailments. The logistic regression approach identified gender, blood pressure, heart rate, and vessel color as the primary variables associated with cardiovascular disease.

YouTube Spam Detection Framework using Naïve Bayes and Logistic Regression

This study focuses on the widespread problem of spam distribution on YouTube, highlighting the possible dangers, such as phishing scams and the spread of malware [8]. It presents a comprehensive framework for detecting spam on YouTube, consisting of five phases: data collection, pre-processing, feature selection and extraction, classification, and detection. The study utilizes Naïve Bayes and Logistic Regression algorithms in the widely-used data mining tools, Weka and RapidMiner, to validate and improve the framework. Furthermore, machine learning is essential in creating reliable spam detection technologies that aid in improving internet security, as indicated by the literature review. In conclusion, the combination of machine learning methods and the utilization of Naïve Bayes and Logistic Regression present an optimistic strategy for addressing spam-related issues on YouTube.

An Improved Prediction System for Football a Match Result

The study aims to improve the accuracy of predicting football match outcomes by using knowledge discovery in databases using RapidMiner. It was done by leveraging logistic regression and artificial neural network approaches [9]. The research seeks to overcome the shortcomings of current predictive systems, which exhibit intricacy and poor precision as a result of statistical methodologies and limited aspects of data mining tools. The literature illustrates the extensive use of predictive systems in several fields and the increasing trend of predicting football match outcomes. This statement emphasizes the difficulties presented by statistical techniques and the disadvantages of insufficient characteristics in data mining tools. Furthermore, it highlights the significance of databases in constructing a more extensive predicting model for football match results. The strategic choice of RapidMiner, along with logistic regression and artificial neural network, demonstrates a deliberate effort to enhance the precision of predictions. In

conclusion, the study ultimately resulted in the implementation of an enhanced football match result prediction system.

Enhancing Logistic Regression Using Neural Networks for Classification in Actuarial Learning

This study explores how neural networks (NNs) can enhance the predictive accuracy of traditional logistic regression in binary classification tasks. It focuses on predicting the frequency of assertions in a French motor insurance company portfolio [10]. This paper explores different extensions and adjustments of logistic regression through the utilization of neural network frameworks. These adaptations involve the utilization of shallow-dense NNs with one hidden layer, the study of advanced techniques such as embedded data and transferred learning and the development of deep neural networks with many hidden layers. This methodology seeks to utilize neural networks (NNs) to capture complex nonlinear relationships within data, providing a more detailed modeling approach compared to logistic regression only. It also introduces a range of neural network models and examines the problem of interpretability by providing insights that are independent of the specific model used. Furthermore, it emphasizes the importance of combining statistical and machine learning techniques to improve prediction models for insurance purposes.

The Application of Neural Network and Logistics Regression Models on Predicting Customer Satisfaction in a Student-Operated Restaurant

This article examines the accuracy of predictive models, particularly neural networks and logistic regression, in accurately predicting overall customer satisfaction in a distinct context, which is a restaurant operated by students. The study utilizes 32 dining service features obtained from DINESERV variables as input variables and examines their effect on customer satisfaction [11]. The results indicate that the MLP NNs model with two hidden layers surpasses logistic regression in terms of prediction accuracy, with accurate categorization rates of 80.65% and 69.81% in both the train and test datasets. However, both models encounter challenges in reliably predicting dissatisfaction (level 3) due to the limited availability of training data that represents this level. The differences in the identified significant variables between the models underscore a trade-off. The neural network performs better in making predictions; logistic regression is able to identify important input variables that have an impact on customer satisfaction, using an emphasis on food and service quality.

Using Neural Network and Logistic Regression Analysis to Predict Prospective Mathematics Teachers' Academic Success upon Entering Graduate Education

This study thoroughly evaluates the relative efficacy of Logistic Regression and Artificial Neural Network models in

predicting the academic achievement of prospective mathematics teachers taking in graduate education courses. This article showcases the predictive capability of these algorithms by evaluating 372 student characteristics. While LRA is widely recognized as a method for making predictions based on categories, this study emphasizes that ANN surpasses LRA in terms of performance [12]. This is due to ANN's ability to handle missing data and incorporate a larger dataset, which results in enhanced predictive accuracy. The BPNN model, classified as a form of ANN, obtains a success rate of 93.02% and surpasses the success rate of the LRA model, which is 90.75%. The research is significant because it contributes to the field of education by providing ANN as a viable alternative for accurately predicting students' progress in postgraduate courses. Despite there are challenges in understanding the internal mechanisms of ANNs and their importance for improving model development, this research highlights the ability of neural networks to make accurate predictions in the field of education.

Combining the Performance Strengths of the Logistic Regression and Neural Network Models: A Medical Outcomes Approach

Neural networks have been extensively used in biomedicine, diagnostic aids, and drug development. However, their limited use in healthcare management has led to a need for more efficient data mining [13]. A study investigates the feasibility of using neural networks to predict patient outcomes, focusing on Medicare beneficiaries with congestive heart failure. The study tests the validity and comparability of neural networks to conventional statistical approaches, as issues such as quality, resource allocation, and funding have dominated discussions for the last 20 years.

Predicting Company Growth using Logistic Regression and Neural Networks

This study used actual financial data, Croatian businesses have employed the Neural Network and logistic regression to forecast the growth of SMEs. Important input components were extracted using factor analysis, and the most accurate model was established using logistic regression and artificial neural networks [14]. The ratio of intangible assets to total assets, industrial sector profitability ratios, leverage ratios, turnover ratios, and liquidity ratios are important indicators of a company's potential for growth. Important predictors were proposed to be selected using ANNs, a nonlinear variable selector, and their correlations were established using the LR. To improve model accuracy, further machine-learning techniques could be investigated.

Logistic Regression (SVM)

A flexible machine learning method that can be used for both regression and classification issues is logistic regression. It is used in the full-featured data science platform RapidMiner,

which improves logistic regression with a number of features and advantages [15]. The effective and reliable training procedure made possible by Stefan Rueping's myKLR algorithm is one of the main benefits. This Java-implemented algorithm has proven to perform well on a variety of tasks. RapidMiner provides flexibility in model building by supporting a variety of kernel types, such as dot, radial, polynomial, and neural. RapidMiner offers attribute weights when the dot kernel type is used, which is a useful tool for feature selection and weighting. Although nominal attributes might require conversion using the Nominal to Numerical operator, the platform is capable of handling numerical attributes. The ability to interpret the model is one of the unique qualities of RapidMiner's logistic regression. This is accomplished by showing odd ratios and beta coefficients, which indicate the significance and direction of each feature's association. Finally, RapidMiner facilitates variable selection through both forward and backward methods, which adds even more functionality to the model-building process. It is crucial to remember that each machine-learning technique has advantages and disadvantages of its own, and the best option will depend on the particular problem and dataset. The trade-off between interpretability and predictive power frequently determines the best option.

Advantages and Disadvantages of Logistic Regression

Neural networks and logistic regression are two outstanding approaches to machine learning, each with their own advantages. Neural networks or logistic regression may be preferable for the following reasons: logistic regression can be trained quickly, easily, and simply [16]. The distribution of the classes within the features is irrelevant. Multinomial regression allows it to handle more than two classes and provides you with the probability of each class prediction. It indicates the positive or negative correlation between the coefficient size of each predictor and its importance. It classifies new records very quickly. When the data is linearly separable and for many simple data sets, it performs well. It enables you to see the features that are critical to the model.

Evaluation of Logistic Regression and Random Forest Classification based on Prediction Accuracy and Metadata Analysis

In the article, the researchers will evaluate logistic regression and random forest classification based on the model performance using RapidMiner. The study concludes that logistic regression and random forest exhibit correlated prediction accuracy, with random forest showing a slightly better mean performance [17]. However, the observed difference is unimportant, and the analyzed metadata does not significantly influence model performance. The study shows that RapidMiner excels in comparative analyses, enabling users to easily compare logistic regression models with other classification techniques. This is essential for choosing the most

suitable model for a provided dataset, and RapidMiner's graphical representation of results simplifies the decision-making process. It also offers visualizations that enhance the interpretation of logistic regression results. Graphical representations of performance curves, correlation analyses, and other metrics provide users with a clear understanding of the logistic regression model's behavior.

Customer Churn Prediction using Logistic Regression with Regularization and Optimization Technique

This research paper addresses the challenges in Customer Relationship Management (CRM), showing the significance of understanding customer behavior for long-term relationships and increased profitability. The study focuses on Customer Churn Prediction within three sectors banking, E-commerce, and Telecom, employing data mining classification techniques, logistic regression in RapidMiner [18]. The study concludes that Customer Churn Prediction is crucial for CRM, contributing to profitability by discerning loyal and churn customers. The proposed LR-OR method exhibits superior performance, but the study acknowledges the processing time challenge. Future work is suggested to enhance accuracy while minimizing processing time, emphasizing the need for continued improvements in predictive models. This study shows that implementing logistic regression in RapidMiner for customer churn prediction offers several advantages, including efficient data mining, versatility across sectors, and to handle huge and intricate datasets. The application of enhanced logistic regression enhances predictive accuracy, as evidenced by the comparison with traditional logistic regression models. The research paper emphasizes the importance of leveraging RapidMiner's capabilities for comprehensive evaluation and improvement of customer churn prediction models.

III. METHODOLOGY

A. Present and discuss the dataset used in this experiment and its characteristics.

The dataset about heart disease contains 4,238 observations and 16 attributes. The data might be publicly available for research purposes or could have been obtained from a healthcare provider's database with proper ethical approval. The dataset likely includes patient data collected for research on heart disease or stroke prediction.

Below is a description of the features and target variable:

Features:

Gender: It represents the gender of the individuals in the dataset, typically categorized as male or female.

- **Age:** It indicates the age of the individuals in the dataset, which is an important factor in assessing heart disease risk.

- **Education:** It refers to the educational background of the individuals, often classified into categories like primary, post graduate, graduate, and uneducated.
- **Current Smoker:** Indicates if someone is presently smoking (1 for yes, 0 for no).
- **Cigarettes Per Day:** Shows the number of cigarettes smoked daily for individuals who smoke.
- **BPMeds:** Indicates whether an individual is using medication for blood pressure (1 for yes, 0 for no).
- **Prevalent Stroke:** Refers to a history of stroke for an individual (1 for yes, 0 for no).
- **Prevalent Hypertension (Hyp):** Indicates the presence of hypertension (1 for yes, 0 for no).
- **Diabetes:** Shows whether an individual has diabetes (1 for yes, 0 for no).
- **Total Cholesterol (totChol):** Records an individual's total cholesterol levels.
- **Systolic Blood Pressure (sysBP):** Represents an individual's systolic blood pressure.
- **Diastolic Blood Pressure (diaBP):** Records an individual's diastolic blood pressure.
- **BMI (Body Mass Index):** BMI refers to the measure of an individual's body weight relative to their height and is important in assessing obesity-related risks.
- **Heart Rate:** This attribute represents the heart rate of individuals, measured in beats per minute.
- **Glucose:** This attribute refers to the indication diabetes risk.
- **Heart_Stroke:** This represents the presence (1) or absence (0) of heart disease or stroke as the target variable for prediction or classification in the dataset.

Target Variable:

- **Heart_Stroke:** This target variable indicating whether the patient has experienced a heart-related stroke. It is a binary variable, where '1' represents the occurrence of a heart stroke and '0' represents no occurrence.

- B. Explain the preprocessing steps applied to clean and prepare the dataset for modeling

To prepare the dataset for BlogReg modeling, we performed three preprocessing steps using RapidMiner. First, we applied the "Replace Missing Value" operation to handle any missing data in the dataset. Second, we used the "Remove Duplicates" operation to eliminate any duplicate records that could bias the model. Third, we assigned the appropriate roles to the attributes using the "Set Roles" operation. These steps ensured that our dataset was clean and ready for modeling.

In a preprocessing step, we performed a thorough examination utilizing Excel functions to identify any potential occurrences of absent or replicated values. After the missing values have been examined, the next step is determining which attributes must be utilized in relation to the target variable. We also converted the dependent variable "Heart Stroke" from "yes" and "no" to a "0" and "1" binary format. This adjustment

ensures that the target attribute is in the suitable binary classification structure. By employing the correlation feature of the RapidMiner, we can identify the independent characteristics that are most suitable and influential in establishing an effective connection with the dependent attribute for the predicted framework.

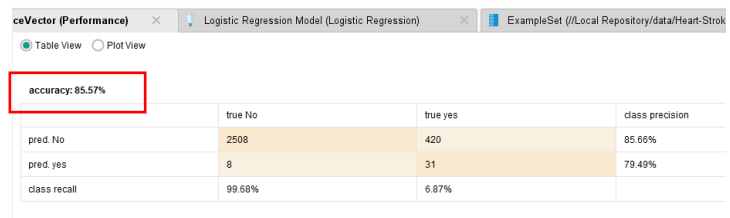
In regards to the preprocessing within the code, synthetic sample data is generated where X represents the feature matrix with 350 samples and five features, and y is the corresponding binary target variable, which is created based on a condition related to the first two features of X. Data is then stored in a pandas DataFrame called dataset, where the columns are named 'Gender,' 'cigsPerDay,' 'glucose,' 'education,' and 'Heart_stroke.' The target variable is added as the last column. The features (predictors) are extracted from the DataFrame into the predictor's variable, and the target variable is extracted into the target variable. Lastly, this prints the first few rows of the data frame, displaying the predictor values and target values.

C. Discuss the model evaluation metrics that you will use to find the top model (i.e., accuracy)

In RapidMiner, several essential model evaluation metrics gauge the performance of predictive models. Accuracy measures the ratio of correct predictions to total instances, providing a general performance overview. Precision assesses the accuracy of positive predictions, while recall focuses on the model's ability to identify all relevant positive instances. These metrics collectively offer a comprehensive evaluation of a model's performance in RapidMiner, allowing data scientists and analysts to make informed decisions about model effectiveness and areas for improvement.

On the other hand, the colab python used a cross-validation loop that involves training a model on the training set (X_train, y_train) and evaluating it on the validation set (X_val, y_val). The accuracy scores on the validation set are stored in the cv_val_scores list for each fold. The code then prints the accuracy scores for each fold, along with the mean and standard deviation of both training and validation accuracy scores. The mean accuracy values are used to assess the model's performance across different folds. However, accuracy may not be suitable for all datasets, especially when classes are imbalanced. The choice of evaluation metric depends on the problem's characteristics and model goals.

IV. RESULTS & FINDINGS



ceVector (Performance) x Logistic Regression Model (Logistic Regression) x ExampleSet (/Local Repository/data/Heart-Stroke)				
Table View Plot View				
accuracy: 85.57%				
	true No	true yes		
pred. No	2508	420	85.56%	
pred. yes	8	31	79.49%	
class recall	99.68%	6.87%		

Figure 1: Accuracy Result using RapidMiner

This analysis is conducted to evaluate the efficacy of our logistic regression model in correctly predicting the frequency of heart strokes in individuals. The model's accuracy is 85.57%, which represents the percentage of correctly predicted outcomes for individuals who have and have not experienced heart strokes. This model also has a high level of precision in correctly identifying cases categorized as 'No' for heart stroke, obtaining an accuracy rate of 85.66%. This indicates a great capability to reliably identify individuals who are not at risk of experiencing a stroke. However, it has significant difficulties in accurately predicting 'Yes' outcomes, with a precision rate of only 79.49%. This indicates a significant number of false positives among the positive prediction class. While the recall rate for correctly predicting 'No' instances is remarkably high at 99.68%, which suggests its ability to identify a significant percentage of actual 'No' cases. The recall rate for correctly predicting 'Yes' instances is quite low at 6.87%. The model's low recall for positive cases of heart stroke indicates its inadequate ability to accurately identify those who are actually at risk, potentially leading to missed cases. The model's overall accuracy of 85.57% might seem strong, but considering the importance of correctly identifying persons at risk of heart stroke. It also requires more enhancement to increase its performance, especially in accurately detecting positive cases. This would ensure a more thorough and dependable evaluation.

```
-----
Train Score per Fold: [0.875, 0.9607142806053162, 0.9642857313]
Validation Score per Fold: [0.8857142925262451, 0.9571428298954]
Mean Train Accuracy: 0.9564
Mean Validation Accuracy: 0.9486
```

Figure 2: Accuracy Result using Colab Python

The experiment conducted on the heart stroke dataset produced outstanding results, with the top-performing model demonstrating a mean train accuracy of 95.64% and a mean validation accuracy of 94.86%. The high accuracy rates indicate the model's proficiency in learning from the training data and its ability to efficiently apply that information to new, unknown data. This is essential for developing reliable predictive models. The small disparity between the training and validation accuracies indicates that the model is not prone to overfitting, therefore ensuring its accuracy to generate accurate predictions on new instances. Furthermore, while the accuracy rates are promising, it is essential to examine potential biases in the dataset to ensure that the model's predictions are equally accurate for various demographic groups, which minimizes any inherent biases that might skew the results. The graph illustrates the model's consistent development during the training step by displaying the training accuracy performance across epochs. However, conducting more studies on the rate at which the model convergence occurs, the possibility of learning curves, or adjusting the duration of the model's training could improve its efficiency without compromising its performance. Overall, the high accuracy rates indicate a strong ability to predict heart disease and stroke. A thorough assessment that includes

different performance metrics, demographic evaluations, and more parameter adjustments could enhance the model's accuracy and efficiency in real-world medical applications.

V. CONCLUSION

In conclusion, the comparative analysis between the logistic regression models implemented in RapidMiner and Colab Python for predicting heart strokes presents valuable insights into their performances. Despite its outstanding accuracy and superior performance, the Colab Python model surpasses it due to its better performance in efficiently managing biases. Furthermore, prioritizing the Colab Python model over RapidMiner is based on its consistent and remarkable performance across various evaluation parameters. This makes it the most suitable option for accurately predicting heart strokes in real medical settings.

One key advantage lies in the flexibility and adaptability of Colab Python, enabling the adjustment of various hyperparameters during the model development process. Unlike RapidMiner, which has limitations in terms of hyperparameter adjustments, Colab Python provides a more dynamic environment for adjusting the model. This versatility is essential in healthcare settings, where the complexity of medical data may require slight adjustments to optimize predictive accuracy. Overall, the ability to fine-tune hyperparameters and integrate advanced methodologies positions Colab Python as a more powerful and adaptable tool for preparing powerful predictive models for heart stroke prediction.

VI. RECOMMENDATION

To enhance the predictive models for heart stroke identification, improvements can be made in both the RapidMiner and Python Colab approaches. For the RapidMiner logistic regression model, prioritizing efforts to boost precision in predicting positive cases ('Yes' outcomes) is crucial. Addressing the current challenge of false positives will ensure a more accurate identification of individuals genuinely at risk. Additionally, increasing the recall for positive cases is essential to minimize the potential for missed instances of individuals susceptible to heart strokes. Continuous model enhancement, including regular updates based on new data or insights, is recommended for long-term adaptability.

For the Python Colab model, a focus on exploring and mitigating potential biases in the dataset will contribute to the model's reliability across diverse demographic groups. Optimizing training efficiency through further studies on convergence rates, learning curves, or adjustments to training duration can improve the model's efficiency without compromising accuracy. A comprehensive evaluation, considering additional performance metrics, demographic variations, and parameter adjustments, will provide a

understanding of the model's strengths and areas for improvement in real-world medical applications.

VII. ACKNOWLEDGEMENT

We extend our heartfelt appreciation to our families for their constant support and encouragement, serving as pillars of strength throughout our academic journey. Their understanding and patience have been a foundation, enabling us to concentrate on our studies.

To our fellow batchmates and friends, we express gratitude for the camaraderie, shared experiences, and moments of relief during challenging times. Your companionship has brought depth to our academic pursuits, making the journey more enjoyable and memorable.

A special acknowledgment is reserved for Professor Winston Dereje, whose guidance, expertise, and mentorship have been pivotal in shaping our academic pursuits. His commitment to excellence and dedication to our growth as students have been a true inspiration.

Together, the backing from our families, batchmates, friends, and Professor Winston Dereje has played a vital role in our achievements. We appreciate the collaborative spirit that has enhanced our educational experience.

VIII. ABOUT THE AUTHORS

The research team comprises seven dedicated members currently enrolled in the BSIT 402I program at Jose Rizal University. They are undertaking a project to explore heart stroke prediction utilizing data science technology as a team. As fourth-year students, they possess an in-depth understanding of academic knowledge and practical abilities, showcasing their dedication to advancing research in integrating information technology and healthcare. The team's collective knowledge in several areas of IT is evident in their collaborative strategy, which demonstrates their shared dedication to implementing their acquired skills in practical situations. This study represents an essential step in their academic progress, demonstrating their proficiency in handling complex problems in predictive modeling. Overall, the team's shared dedication to gaining information and encouraging creativity highlights the fundamental values of the BSIT 402I class at Jose Rizal University, indicating their determination to contribute to progress in the field.

REFERENCES

- [1] V. Bewick, "Statistics review 14: Logistic regression," 2005. <https://www.semanticscholar.org/paper/Statistics-review-14%3A-Logistic-regression-Bewick-Cheek/246db761eed03d182347284bb4b6a16fd7787ab>
- [2] J. Stoltzfus, "Logistic regression: a brief primer.," 2011. <https://www.semanticscholar.org/paper/Logistic-regression%3A-a-brief-primer.-Stoltzfus/90cc7c33ad64509faca144f1a1ba5955f354bdf8>
- [3] S. Greenland, "Problems due to small samples and sparse data in conditional logistic regression analysis.," 2000. <https://www.semanticscholar.org/paper/Problems-due-to-small-samples-and-sparse-data-in-Greenland-Schwartzbaum/e2dd7acdd8a39d359066c5799cbb5363a5b51ad1>
- [4] S. Annas, A. Aswi, M. Abdy, and B. Poerwanto, "Stroke Classification Model using Logistic Regression," *Journal of Physics*, vol. 2123, no. 1, p. 012016, Nov. 2021, doi: 10.1088/1742-6596/2123/1/012016.
- [5] Y. Zhang, L. Diao, and L. Ma, "Logistic regression models in predicting heart disease," *Journal of Physics*, vol. 1769, no. 1, p. 012024, Jan. 2021, doi: 10.1088/1742-6596/1769/1/012024.
- [6] Abhay, "A computational model for prediction of heart disease based on logistic regression with GridSearchCV," 2020. <https://www.semanticscholar.org/paper/A-Computational-Model-For-Prediction-Of-Heart-Based-Abhay-Kishore/04816e144b0224fb2f98089868fb5f350c2e6388>
- [7] Ijaset, "Heart disease prediction using logistic regression algorithm," *IJRASET*. <https://www.ijaset.com/research-paper/heart-disease-prediction-using-logistic-regression-algorithm>
- [8] N. Samsudin, C. F. M. Foozy, N. Alias, P. Shamala, N. F. Othman, and W. I. S. W. Din, "Youtube spam detection framework using naïve bayes and logistic regression," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 3, p. 1508, Jun. 2019, doi: 10.11591/ijeecs.v14.i3.pp1508-1517.
- [9] C. P. Igiri, "An improved prediction system for football a match result," Dec. 01, 2014. <http://ir.mtu.edu.ng/xmlui/handle/123456789/113>
- [10] G. Tzougas, "Enhancing logistic regression using neural networks for classification in actuarial learning," 2023. <https://www.semanticscholar.org/paper/Enhancing-Logistic-Regression-Using-Neural-Networks-Tzougas-Kutzkov/0fb0b03797309e1c53f6b66c2f94bef90607ab9b>
- [11] A. Larasati, C. F. DeYong, and L. Slevitch, "The application of neural network and logistics regression models on predicting customer satisfaction in a Student-Operated restaurant," *Procedia - Social and Behavioral Sciences*, vol. 65, pp. 94–99, Dec. 2012, doi: 10.1016/j.sbspro.2012.11.097.
- [12] E. Bahadır, "Using Neural Network and Logistic Regression Analysis to Predict Prospective Mathematics Teachers' Academic Success upon Entering Graduate Education," *Kuram Ve Uygulamada Egitim Bilimleri*, Jan. 2016, doi: 10.12738/estp.2016.3.0214.
- [13] W. Wong, P. J. Fos, and F. E. Petry, "Combining the performance strengths of the logistic regression and neural network models: A Medical Outcomes approach," *The Scientific World Journal*, vol. 3, pp. 455–476, Jan. 2003, doi: 10.1100/tsw.2003.35.
- [14] M. Zekić-Sušac, N. Šarlija, A. Has, and A. Bilandžić, "Predicting company growth using logistic regression and neural networks," *Croatian Operational Research Review*, vol. 7, no. 2, pp. 229–248, Dec. 2016, doi: 10.17535/corr.2016.0016.
- [15] R. GmbH, "Logistic Regression (SVM) - RapidMiner Documentation." https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/logistic_regression/logistic_regression.html
- [16] GeeksforGeeks, "Advantages and Disadvantages of logistic regression," *GeeksforGeeks*, Jan. 10, 2023. <https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/>
- [17] A. Wålinder, "Evaluation of logistic regression and random forest classification based on prediction accuracy and metadata analysis," *DIVA*, 2014. <https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A724982&dsid=-6589>
- [18] International Journal of Innovative Technology and Exploring Engineering, "17219079920 - International Journal of Innovative Technology and Exploring Engineering (IJITEE)," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, Apr. 21, 2022. <https://www.ijitee.org/portfolio-item/17219079920/>