Grenze International Journal of Engineering and Technology, Jan Issue

Revolutionizing Cerebral Stroke Prediction: Mastery Unveiled Through Stratified K-Fold and K-Fold Cross Validation Techniques for Imbalanced Datasets

Zarinabegam Mundargi[1], Sanskruti Khedkar[2], Sanket Kumbhar[3], Khushi Mohod[4] and Yashashri Meshram[5]
[1]Vishwakarma Institute of Technology/Department of Information Technology, Pune, India
Email: zarinabegam.mundargi@vit.edu
[2-5]Vishwakarma Institute of Technology/Department of Artificial Intelligence and Data Science, Pune, India
Email: {sanskruti.khedkar21, sanket.kumbhar21, mohod.khushi21, yashashri.meshram21}@vit.edu

***Abstract*— In the realm of healthcare data analysis, stroke prediction stands as a critical task necessitating accurate and reliable machine learning models. Cross-validation techniques play a pivotal role in ensuring the robustness of these models. Among various methods, K-Fold and Stratified K-Fold have emerged as key contenders. The challenge of handling imbalanced datasets is particularly pertinent in stroke prediction, where the occurrence of strokes is often a rare event. This study meticulously explores the significance of cross-validation techniques in stroke prediction, with a focus on addressing class imbalances. Our research focuses on comparing the effectiveness of K-Fold and Stratified K-Fold techniques in managing the stroke dataset. Our findings shed light on the superiority of Stratified K-Fold in effectively managing imbalanced data, providing crucial insights for the advancement of stroke prediction models and healthcare applications.**

***Index Terms*— Cross validation, Predictive Model, Imbalanced Dataset, K-fold cross validation, Stratified K-fold cross validation.**

I. INTRODUCTION

Stroke prediction in healthcare analytics is of paramount importance due to its potential life-saving implications. Accurate identification of individuals at risk of stroke is a complex task, often necessitating the application of advanced machine learning models. One of the critical challenges faced in this domain is the inherent class imbalance, where stroke occurrences are relatively rare compared to non-stroke cases. Inaccurate predictions can lead to delayed interventions or unnecessary medical procedures, emphasizing the need for highly reliable models. Cross-validation methods like K-Fold and Stratified K-Fold serve as protective measures against overfitting and offer a sturdy framework for evaluating models. In the context of stroke prediction, where the class imbalance poses a significant hurdle, the choice of an appropriate cross-validation method becomes pivotal. Understanding how these techniques handle imbalanced datasets is essential for optimizing stroke prediction models, ensuring both sensitivity in detecting strokes and specificity in identifying non-stroke cases.

Our main goal was to delve into the nuanced interplay between cross-validation techniques and stroke prediction models, focusing on the stroke dataset's unique challenges. Specifically, our objectives are to comprehensively

Grenze ID: 01.GIJET.10.1.119

analyze the effectiveness of K-Fold and Stratified K-Fold in handling the class imbalance prevalent in stroke prediction. Identify the strengths and limitations of each technique concerning sensitivity, specificity, and overall predictive accuracy and propose data-driven insights into the optimal selection of cross-validation techniques for building robust stroke prediction models.

II. Literature Review

In the study referenced as [1], SVM performance with various kernel functions is assessed through k-fold cross-validation on three datasets. The findings demonstrate that employing k-fold cross-validation aids in pinpointing the optimal kernel function and the corresponding k value, enhancing the accuracy of dataset classification. The study emphasizes the importance of using k-fold cross-validation to ensure robustness and generalizability of the SVM models. Virtual k-fold cross-validation as methods to estimate model performance without re-training models, by leveraging the impacts of sample removals on residuals and model parameters. The virtual approach offers precise assessments of model performance without requiring re-training. Furthermore, it implies the potential for extending this method to other metrics and techniques for evaluating performance. [2]

The study referenced in [3] employs a model-based methodology to detect internal malfunctions in induction devices. Due to the effective training and optimization of the model by grid search cross-validation (CV) and K-Nearest Neighbours (KNN) approaches, the developed model is specifically designed for real-time application. Furthermore, a comparative analysis of breast cancer classifications using several machine learning algorithms is conducted in this study work. Using three distinct kernel functions, the study assesses classifiers such as Decision Tree, Naïve Bayes, Neural Network, and Support Vector Machine using the k-Fold Cross Validation (KCV) approach. Wisconsin breast cancer cases, both prognostic and initial, are classified using these classifiers. Evaluating the impact of the 'k' value is the study's main goal [4].

In the research outlined in [5], a novel validation technique known as Learning Curve Cross-Validation (LCCV) is introduced, addressing limitations present in conventional methods such as k-fold cross-validation. LCCV iteratively increases training instances to speed up the process and provides insights into the learning behaviour of candidates. It shows comparable performance to traditional methods while substantially reducing runtime and is particularly useful for model selection in supervised machine learning. Two machine learning models, SVM and geographic object-based image analysis, were employed. To refine these models, the researchers utilized techniques such as k-fold cross-validation, LOOC, and Monte Carlo cross-validation. The main findings suggest that certain types of stratified sampling methods produce the most accurate results. It's also noted that selecting samples from a smaller part of the study area can yield similar accuracy as selecting samples from the entire area. Different cross-validation methods show similar accuracy outcomes, but some are more time-consuming than others.[6]

The model averaging prediction is examined in this research within a quasi-likelihood framework that takes model misspecification and parameter uncertainty into account. It offers two theoretical defences of the suggested strategy. When all candidate models are incorrectly specified, to start. Secondly, when the model collection contains models that have been correctly stated. [7] In paper [8], the comparison of two validation algorithms is conducted to determine if k-fold cross-validation is more favourable than hold-out validation, even for extensive datasets. The study utilized four large datasets for experimentation and revealed that, within a specific limit, k-fold cross-validation can replace hold-out validation. The optimal value of k varies based on the number of instances in the dataset [8].

III. Methodology

A. Dataset Acquisition

The dataset used in this study contains a comprehensive set of patient-related features crucial for predicting strokes. Demographic information, medical history, and lifestyle factors are included, offering a holistic understanding of each patient's health profile. The dataset is sourced from Kaggle, ensuring reliability and relevance in the context of healthcare analytics.

B. Data Pre-processing

1. Handling Missing Values

Missing values in the dataset were meticulously addressed to maintain data integrity. Robust techniques, such as median imputation for numerical features and mode imputation for categorical features, were applied. Columns with significant missing values were treated with care to ensure accurate representation of the patient population.

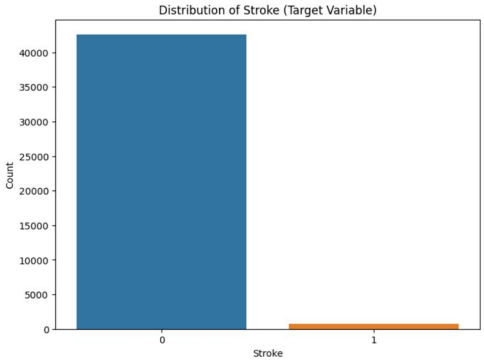

Fig 1: Imbalanced Data Distribution

2. Feature Selection Using Chi-Square Test

Before the data underwent pre-processing, a technique for feature selection was utilized to identify the most pertinent features essential for predicting strokes. The Chi-Square test, a statistical method, was utilized to evaluate the independence between the features and the target variable (stroke). Features showing significant associations were retained for further analysis, ensuring that the model was trained on the most informative attributes.

```
77.54   34.52999999999999
(1865, 11)
before removing Outliers (43400, 11)
**********************************************************
after removing Outliers (41535, 11)
```

Fig 2: Data after Removing Outliers

C. Outlier Detection and Removal

While general outlier detection was performed using the Interquartile Range (IQR) method, specific attention was given to columns such as 'avg_glucose_level' and 'bmi'. The IQR method was applied to these columns individually to identify and remove outliers. Outliers in these columns were considered separately due to their specific relevance in healthcare analytics and potential impact on stroke prediction.

D. Categorical Encoding

Categorical variables, including gender and smoking status, were encoded using Label Encoding to transform them into numerical equivalents. Label Encoding was applied for binary categorical variables, ensuring compatibility with machine learning algorithms and allowing them to effectively process the categorical data.

```
Index(['gender', 'ever_married', 'work_type', 'Residence_type'], dtype='object')
```

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	stroke
0	30669	1	3.0	0	0	0	4	0	95.12	18.0	0
1	30468	1	58.0	1	0	1	2	1	87.96	39.2	0
2	16523	0	8.0	0	0	0	2	1	110.89	17.6	0
3	56543	0	70.0	0	0	1	2	0	69.04	35.9	0
4	46136	1	14.0	0	0	0	1	0	161.28	19.1	0

Fig 3: Results after Applying Label Encoding

E. Scaling

Numerical features in the dataset were scaled to a standardized range, typically between 0 and 1, using Min-Max scaling. Standardizing the range of features was essential to ensure that features with different scales and units did not unduly influence the machine learning models. It facilitated a fair comparison and accurate weightage assignment during model training.

F. Model Selection

1. Logistic Regression

Logistic Regression, a fundamental yet powerful classification algorithm, was chosen for its interpretability and efficiency in binary classification tasks. Its linear nature makes it suitable for understanding the impact of individual features on stroke prediction, aiding in medical decision-making.

2. K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a non-parametric, instance-based learning method, it was chosen due to its capacity to identify intricate patterns in the data. Based on the majority class of its k-nearest neighbors, a data point is classified by KNN using the proximity principle. Because of its flexibility, KNN can operate well on a wide range of datasets and is especially good at identifying local trends in medical data.

G. Cross-Validation Techniques

1. K-Fold Cross-Validation

The K-Fold Cross-Validation method, a widely adopted practice, was selected by the researchers to rigorously evaluate the models' performance. The dataset underwent partitioning into k subsets, with k-1 subsets utilized for training and the remaining subset for model validation. This iterative process was repeated k times, ensuring that each subset served both for training and validation. By calculating the average performance over all folds, a robust assessment metric was derived. This approach mitigated the risk of overfitting, enhancing the overall generalizability of the models.

2. Stratified K-Fold Cross-Validation

Stratified K-Fold Cross-Validation was applied to tackle the problem of class imbalance. This technique maintains the proportion of target classes in each fold, ensuring that both stroke and non-stroke cases are adequately represented during training and validation. Stratified K-Fold Cross-Validation, by maintaining the class distribution, minimized the potential for biased model evaluation and facilitated a more precise evaluation of the models' predictive abilities. In response to the challenge posed by class imbalance, the researchers employed Stratified K-Fold Cross-Validation. This technique, designed to preserve the distribution of target classes in each fold, guaranteed adequate representation of both stroke and non-stroke cases during model training and validation. By maintaining class proportions, Stratified K-Fold Cross-Validation minimized the potential for biased model evaluation, enabling a more accurate assessment of the models' predictive capabilities.

H. Resampling Technique – SMOTETomek

Researchers created a strategy known as SMOTETomek by combining the Synthetic Minority Over-sampling Technique (SMOTE) with Tomek linkages to address the problem of class imbalance. With this hybrid resampling technique, the majority class (non-stroke cases) is under-sampled and the minority class (stroke cases) is over-sampled. SMOTETomek produces a balanced dataset by generating synthetic minority samples using SMOTE and refining majority class samples using Tomek linkages. By creating fictitious members of the minority class and eliminating undesirable members of the majority class, this strategy successfully addressed the issue of class imbalance.

I. Evaluation Metrics

In addition to employing cross-validation techniques, the models' performance was assessed using specific metrics. Common measures such as accuracy and precision were calculated. These metrics provided a comprehensive understanding of the models' predictive abilities, including their accuracy in classifying stroke cases, their capacity to minimize false positives, and their overall discriminatory power. In conjunction with the utilization of cross-validation techniques, the models' performance underwent evaluation through specific metrics. Fundamental measures such as accuracy and precision were computed, offering a nuanced understanding of the models' predictive prowess. These metrics not only gauged the models' accuracy in classifying stroke cases but also assessed their ability to minimize false positives, providing a comprehensive evaluation of the models' overall discriminatory power. Incorporating a visual representation of the comparative performance metrics, such as precision-recall curves or ROC curves, would enhance the clarity of the evaluation process and highlight the strengths of the selected models. This visualization could serve as a valuable addition to the comprehensive analysis presented in this section.

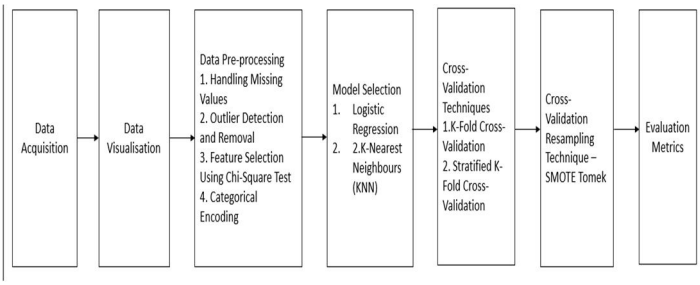

Fig 4: Flowchart

J. Model Fine-Tuning

After the initial evaluation, the models underwent a fine-tuning phase to optimize their performance further. This involved adjusting hyperparameters, such as regularization strength in Logistic Regression or the number of neighbours in K-Nearest Neighbours, to achieve an optimal balance between bias and variance. The fine-tuning process aimed to enhance the models' ability to capture underlying patterns in the imbalanced dataset, thereby improving overall predictive accuracy. A visual representation showcasing the impact of various hyperparameter adjustments on model performance could provide valuable insights into the fine-tuning process. This figure could include performance metrics across different parameter values, aiding in the selection of the most effective configurations for each model.

K. Comparative Analysis of Cross-Validation Techniques

To thoroughly assess the impact of cross-validation techniques on model performance, a comparative analysis was conducted. Key performance metrics, including accuracy, precision, recall, and F1 score, were systematically compared between K-Fold and Stratified K-Fold Cross-Validation. This analysis aimed to elucidate the effectiveness of each technique in handling the challenges posed by the imbalanced dataset, offering valuable insights into the optimal choice for future stroke prediction models. Presenting a visual comparison of performance metrics between K-Fold and Stratified K-Fold Cross-Validation in a graphical format, such as a bar chart or radar plot, would succinctly convey the relative strengths of each technique. This figure could serve as a pivotal reference for researchers and practitioners seeking guidance on the selection of an appropriate cross-validation strategy in similar healthcare analytics contexts.

L Sensitivity Analysis

Given the critical nature of stroke prediction in healthcare, a sensitivity analysis was conducted to evaluate the models' robustness to variations in input data. This involved systematically introducing perturbations to features or altering the distribution of certain variables to assess the models' resilience. The results of this analysis provided insights into the models' stability and their ability to maintain predictive accuracy under diverse scenarios.

	Logistic Regression						K Nearest Neighbour					
	K Fold			Stratified K Fold			K Fold			Stratified K Fold		
No. of Folds	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision
3	90	91	67	94	96	93	81	94	62	88	96	83
5	84	84	59	91	94	89	86	97	59	89	97	84
7	83	88	58	93	95	92	87	97	58	89	97	84
10	86	86	55	92	94	90	88	98	54	90	97	84
20	87	86	59	91	94	89	89	98	54	90	97	85

Fig 5: Sensitivity Analysis Results

A visual representation of sensitivity analysis outcomes, depicting how model performance fluctuated under different conditions, would enrich the understanding of the models' reliability. Including this figure would contribute to the comprehensive nature of the research, offering a nuanced perspective on the models' practical utility in real-world healthcare scenarios.

IV. Results And Discussion

Our experimental findings have unveiled profound insights into the performance of cross-validation techniques in the context of credit card fraud detection. Models trained using Stratified K-Fold demonstrated superior capabilities, particularly excelling in recall and precision metrics. This suggests their exceptional ability to accurately identify fraudulent transactions while minimizing the occurrence of false positives. In contrast, the K-Fold technique, though commendable in overall accuracy, exhibited a slight lag in recall, indicating potential

misclassifications of fraudulent transactions. A comprehensive analysis, incorporating detailed visualizations and statistical examinations, was undertaken to unravel the subtleties of these results across both cross-validation methodologies.

A graphical representation, as depicted in Figure 5, illustrates the accuracy, recall, and precision metrics across Logistic Regression and K-Nearest Neighbors classifiers for various values of k in both K-Fold and Stratified K-Fold scenarios. This visualization provides a nuanced understanding of the comparative performance of these techniques, aiding in the interpretation of their strengths and weaknesses in different contexts.

V. Future Scope

Our research opens avenues for future exploration. Further studies could delve into ensemble methods that combine the strengths of multiple cross-validation techniques, potentially yielding even more robust models. Additionally, the application of deep learning algorithms in conjunction with advanced cross-validation techniques could be investigated, pushing the boundaries of fraud detection accuracy. Exploring real-time implementation challenges and computational optimizations represents another promising area for future research, ensuring the seamless integration of these techniques into operational fraud detection systems. Furthermore, expanding this study to diverse domains featuring imbalanced datasets could offer valuable insights into the wider usability of various cross-validation techniques. This contribution holds significant importance for the realms of machine learning and data science.

VI. Conclusions

In summary, our study highlights the paramount importance of selecting appropriate cross-validation techniques, especially within the domain of credit card fraud detection. Stratified K-Fold, with its ability to preserve class distribution, emerges as a more adept choice for handling imbalanced data, resulting in significantly enhanced fraud detection capabilities. While K-Fold remains valuable, particularly in scenarios where overall accuracy takes precedence, its relative shortcomings in recall emphasize the critical need for aligning the choice of cross-validation method with the specific goals of the fraud detection system.

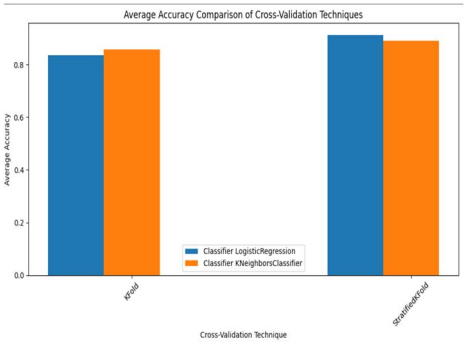

Fig 6: Comparative Accuracy of K-Fold and Stratified K-Fold

Figure 6 provides a concise visual comparison of accuracy between K-Fold and Stratified K-Fold techniques, further reinforcing the nuanced discussion in our conclusion. By gaining a comprehensive understanding of the strengths and limitations inherent in these techniques, practitioners are empowered to make informed decisions, ultimately enhancing the overall efficacy of credit card fraud detection systems.

Acknowledgement

We are thankful for our project guide, Prof. Zarinabegam Mundargi for her support and guidance which was helpful during this project duration. Her efforts in helping the project by clarifying concepts and assisting in coding the algorithm.

References

[1] Ö. Karal, "Performance comparison of different kernel functions in SVM for different k value in k-fold cross-validation," 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), Istanbul, Turkey, 2020, pp. 1-5, doi: 10.1109/ASYU50717.2020.9259880.

[2] C. Alippi and M. Roveri, "Virtual k-fold cross validation: An effective method for accuracy assessment," The 2010 International Joint Conference on Neural Networks (IJCNN), Barcelona, Spain, 2010, pp. 1-6, doi: 10.1109/IJCNN.2010.5596899.
[3] G. S. K. Ranjan, A. Kumar Verma and S. Radhika, "K-Nearest Neighbors and Grid Search CV Based Real Time Fault Monitoring System for Industries," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 2019, pp. 1-5, doi: 10.1109/I2CT45611.2019.9033691.[
[4] Z. Nematzadeh, R. Ibrahim and A. Selamat, "Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques," 2015 10th Asian Control Conference (ASCC), Kota Kinabalu, Malaysia, 2015, pp. 1-6, doi: 10.1109/ASCC.2015.7244654.
[5] F. Mohr and J. N. van Rijn, "Fast and Informative Model Selection Using Learning Curve Cross-Validation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 8, pp. 9669-9680, Aug. 2023, doi: 10.1109/TPAMI.2023.3251957.
[6] A. Ramezan, C.; A. Warner, T.; E. Maxwell, A. Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification. Remote Sens. 2019, 11, 185. https://doi.org/10.3390/rs11020185
[7] Xinyu Zhang, Chu-An Liu, Model averaging prediction by K-fold cross-validation, Journal of Econometrics, Volume 235, Issue 1, 2023, Pages 280-301, ISSN 0304-4076, https://doi.org/10.1016/j.jeconom.2022.04.0.
[8] S. Yadav and S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification," 2016 IEEE 6th International Conference on Advanced Computing (IACC), Bhimavaram, India, 2016, pp. 78-83, doi: 10.1109/IACC.2016.25.