Comprehensive Evaluation of Regression and Classification Models on Brain Stroke Datasets

Dimitar Trajkov, Ana Kostovska, Panče Panov, Dragi Kocev
Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
dimitar.trajkovv@gmail.com,ana.kostovska@ijs.si,pance.panov@ijs.si,dragi.kocev@ijs.si

ABSTRACT

This paper investigates the application of machine learning models for predicting brain stroke outcomes, leveraging publicly available datasets. We evaluate the performance of various classification and regression models, including ensemble methods such as AdaBoost, Gradient Boosting, and Random Forest, across eight datasets related to stroke prediction. Our results show that data quality and dataset characteristics have a more significant impact on model performance than the choice of algorithm, underscoring the importance of high-quality, well-curated data in achieving accurate and reliable predictions. Additionally, we emphasize the need for transparency, reproducibility, and traceability in AI research, highlighting the challenges associated with the scarcity of publicly available stroke datasets. This study provides a foundation for developing more trustworthy AI tools for stroke prediction and encourages further efforts in data sharing and model validation.

KEYWORDS

Brain stroke, scientific benchmarking study, stroke outcome prediction, data quality in AI, AI transparency and reproducibility

1 INTRODUCTION

Brain stroke is a significant global health challenge, ranking as one of the leading causes of mortality and long-term disability. The World Stroke Organization–Lancet Neurology Commission Stroke Collaboration Group [4] has projected that the mortality will increase from 6.6 million people worldwide in 2020 up to 9.7 million in 2050. Beyond the mortality statistics, brain stroke leaves survivors with debilitating effects (with disability-adjusted life years rising from 144.8 millions to 189.3 millions), severely impacting their quality of life. The ability to accurately predict and prevent brain strokes through accessible and straightforward measures can revolutionize public health strategies, especially in low- and middle-income regions where healthcare resources are often limited and the burden of stroke is most pronounced.

In today's data-driven era, the *scarcity of publicly available clinical datasets on brain stroke* presents a critical barrier to advancing research and developing effective predictive models. Hospitals and medical institutions, governed by privacy regulations and the imperative to protect patient confidentiality, are often hesitant to share datasets, even in anonymized forms. Therefore, even the few publicly available datasets are from unknown and unverified sources with no possibility to check their validity.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

Brain stroke [14] is influenced by both non-modifiable factors, such as age, genetic predisposition, and gender, with men generally at higher risk and women more vulnerable during pregnancy and postpartum, and modifiable factors that can be managed through lifestyle changes and medical interventions. Key modifiable risk factors include hypertension, high cholesterol, diabetes, obesity, smoking, atrial fibrillation, and heart-related issues, which can lead to ischemic strokes. Physical inactivity, excessive alcohol consumption, and poor diet further elevate stroke risk, making prevention through lifestyle modification essential for reducing the overall stroke burden.

The need for the use of AI in analyzing brain stroke data is highlighted by its ability to handle the complexity and volume of medical data, including clinical and imaging data, that traditional methods cannot efficiently process [14, 1, 13, 4, 11]. AI models, particularly machine learning (ML), are being used to predict stroke outcomes by processing large datasets with precision, which can help clinicians make more informed decisions. AI aids in diagnosing and predicting the progression of stroke, improving treatment response predictions, and supporting early interventions that are crucial for stroke recovery and prevention.

AI-driven predictive models have been designed to learn from stroke data to forecast outcomes such as mortality, functional impairment, and recovery potential. ML models like support vector machines, random forests, and neural networks have been employed to predict key outcomes using structured clinical data. These models not only provide personalized prognoses but also have the potential to improve patient care by identifying high-risk individuals early. However, challenges remain in integrating these models into clinical practice due to issues like small datasets and poor reporting standards in existing studies.

For AI to become a trustworthy resource in stroke care, transparency, reproducibility, and traceability are essential. There is a growing demand for the reproducibility of AI-based research, which is necessary to ensure that models can be independently validated and applied to different patient populations. In this work, we are making the first step towards providing such trustworthy resources for brain stroke data.

2 DATA AND METHOD DESCRIPTION

In our study, we collected a total of 8 publicly available (tabular) datasets related to brain stroke: four regression datasets and four classification datasets. Of the classification datasets, two are binary classification datasets, and two address multi-class classification problems. Five of the datasets were found at the repository `Data.World`, and 3 at the repository `Kaggle`. Table 1 provides an overview of the datasets used in this study. It includes the names of the datasets, the number of instances, the number of features, and specifies whether each dataset is used for a classification (C) or regression (R) task.

We evaluated the performance of a broad spectrum of models implemented in the `scikit-learn` toolbox [10] to explore

Table 1: Datasets used in the study with hyperlinks, number of instances, features, and task type regression (R) and classification (C).

Dataset Name	Num. of Instances	Num. of Features	Task
Ischemic Stroke 30-Day Mortality and 30-Day Readmission Rates[5]	2188	10	R
Stockport Local Health Characteristics[2]	190	18	R
All Payer In-Hospital/30-Day Acute Stroke Mortality Rates by Hospital (SPARCS)[6]	137	14	R
Brain Stroke Dataset[8]	600	9	C
Brain stroke prediction dataset[9]	4981	11	C
Cerebral Stroke Prediction-Imbalanced Dataset[7]	43400	12	C
Mortality from Stroke[3]	231	9	R
Prognostication of Recovery from Acute Stroke (PRAS Dataset)[12]	161	110	C

different approaches to prediction and analysis. For the classification datsets, we utilized the following different methods. First, we used ensemble methods, such as **AdaBoostClassifier**, **BaggingClassifier**, **RandomForestClassifier**, **GradientBoostingClassifier**, **XGBClassifier** (from the `XGBoost` library), and **LightGBMClassifier**, for their ability to improve predictive accuracy by combining multiple weak learners. These models are particularly effective in capturing complex, non-linear relationships in the data. We also incorporated linear models like **LogisticRegression**, which are valued for their interpretability and simplicity. Other classifiers included **DecisionTreeClassifier**, **KNeighborsClassifier**, **MLPClassifier**, **QuadraticDiscriminantAnalysis**, **RadiusNeighborsClassifier**, **SGDClassifier**, and **SupportVectorClassifier (SVC)**, each contributing unique strengths to the classification tasks.

For the regression datasets, we also evaluated a variety of models. Similarly as for the classification datasets, we used different ensemble methods such as **AdaBoostRegressor**, **BaggingRegressor**, **RandomForestRegressor**, **GradientBoostingRegressor**, **HistGradientBoostingRegressor**, **LightGBMRegressor**, and **XGBoostRegressor** (from the `XGBoost` library). Linear models, including **LinearRegression**, **RidgeRegression**, **LassoRegression**, **LassoLars**, **ElasticNetRegression**, **BayesianRidgeRegression**, **TheilSenRegressor**, **HuberRegressor**, **RAN-SACRegressor**, **PassiveAggressiveRegressor**, **SGDRegressor**, **LeastAngleRegression**, and **OrthogonalMatchingPursuit**, were employed for their simplicity and effectiveness in datasets with linear relationships. Additionally, **GaussianProcessRegressor** and **KNeighborsRegressor** were included to capture local data structures and model complex relationships, while **MLPRegressor** was used for its deep learning capabilities. Finally, we explored the performance some specific regressors such as **OrdinalRegression** (from the `mord` library) and **TweedieRegressor**.

3 DESIGN OF THE EXPERIMENTAL STUDY

Figure 1 illustrates the design of the executed experimental study. After identification and categorization of relevant datasets and separating them into regression and classification tasks based on the target variable, we manually examined each dataset to identify those that required manual preprocessing. The preprocessing steps included several standard procedures applied across all datasets: removal of features with constant values for all examples or missing values for more than 70% of the examples, removal of identifiers, standardization of the numeric features (to mean values zero with standard deviation of one), one-hot encoding for nominal features, and mapping of values for the ordinal features.

Following the data preprocessing, we executed an exhaustive grid search across a broad spectrum of hyperparameter values, using nested 3-cross-validation to select the optimal parameter configurations (using the mean squared error for the regression datasets, and the F1 score for the classification datasets). Nested cross-validation was chosen for its ability to provide an unbiased evaluation of the model's performance by incorporating both an inner loop (3-fold) for hyperparameter tuning and an outer loop (10-fold) for model evaluation. The performance of the models was assessed using a variety of evaluation measures such as accuracy, balanced accuracy, precision, average precision, recall, F1 score, jaccard score, fowlkes mallows score, cohen kappa score, matthews correlation coefictien and others for clasification tasks and mean absolute error, mean squared error, median absolute error, mean percentage error, relative squared error, theil's u statistic and much more for the regression tasks.

Furthermore, to facilitate deeper insights into the data and model performance, we calculated a variety of meta-features describing the datasets. These included basic features such as the number of instances, features, and the proportion of numeric, nominal, binary, and constant features, then also statistical meta-features like geometric, harmonic, and arithmetic means, median, standard deviation, as well as theoretical meta-features such as entropy, correlation, principal component analysis (PCA), and mutual information and more.

All of the information about the experimental procedures and the specific experiments on the datasets using the selected methods are diligently documented in a JSON file. This facilitates traceability and reproducibility of the executed experiments.

4 RESULTS AND DISCUSSION

Table 2 lists the performance of all models on the classification datasets, as measured by the F1 score. The overall impression is that the obtained performances are comparable, with only marginal differences observed. Ensemble models generally performed slightly better, with AdaBoost and Gradient Boosting leading the way in terms of F1 score. Conversely, K-Nearest Neighbors (KNN) showed the lowest performance in this regard. In addition to the F1 score, other evaluation metrics exhibit similar patterns, highlighting their high correlation with each other (as illustrated in Figure 3). This correlation suggests that if a model excels in one metric, it is likely to perform consistently well across other metrics as well. Figure 2 presents a violin plot of the F1 scores evaluated on the test data from the **Brain Stroke Dataset** [8], providing a visual representation of the distribution and variability in model performance.

Table 3 presents the results obtained for the regression tasks using the Root Mean Squared Error (RMSE). We can observe that

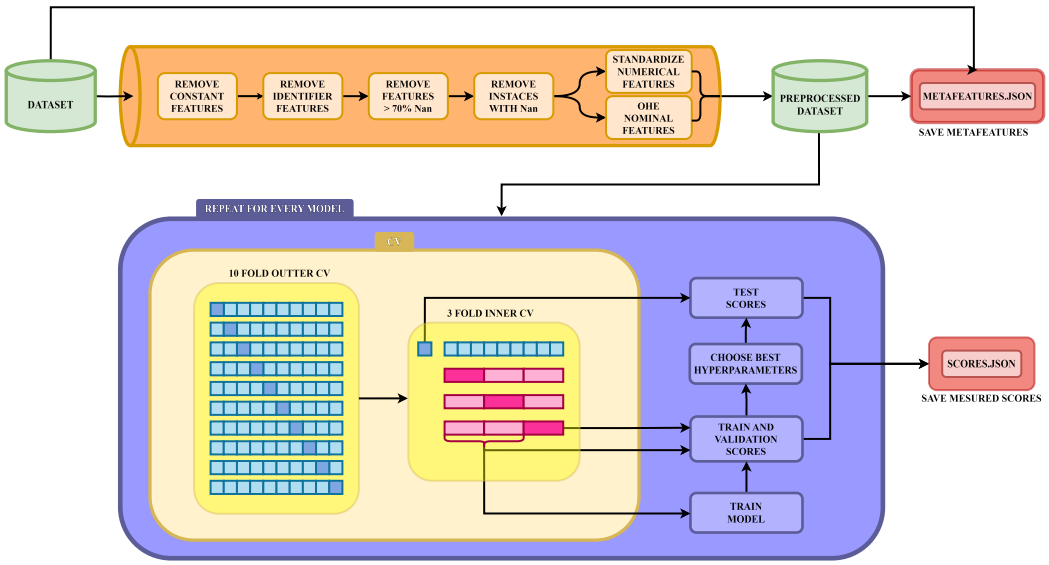

Figure 1: An overview of the procedures used to execute the experimental study including the preprocessing steps, hyperparameter optimization with nested cross-validation, and the calculation of the meta-features of the datasets.

Dataset		AdaBoost	Bagging	Decision Tree	Gaussian distribution	Gradient Boosting	KNN	Logistic Regression	Multi-layer Perceptron	Quadratic Discriminant Analysis	Random Forest	SGD
Brain Stroke Dataset	Mean	1.000	0.986	0.736	0.470	1.000	0.395	0.773	0.732	0.430	1.000	0.951
	Std	0.000	0.042	0.127	0.117	0.000	0.130	0.121	0.093	0.090	0.000	0.101
Brain stroke predictic..	Mean	0.347	0.340	0.354	0.324	0.339	0.332	0.349	0.347	0.301	0.324	0.359
	Std	0.053	0.042	0.049	0.036	0.048	0.037	0.048	0.047	0.045	0.041	0.057
Cerebral Stroke Pred..	Mean	0.246	0.229	0.207	0.172	0.248	0.096	0.233	0.129	0.129	0.252	0.215
	Std	0.074	0.041	0.054	0.038	0.073	0.026	0.089	0.066	0.066	0.073	0.097
Prognostication of..	Mean	0.097	0.089	0.090	0.037	0.101	0.048	0.080	0.016	0.030	0.105	0.078
	Std	0.018	0.014	0.017	0.007	0.018	0.014	0.017	0.019	0.010	0.023	0.020

Table 2: Mean and standard deviation of F1 scores for each model across classification datasets.

Dataset		AdaBoost	Bayesian Ridge	Decision Tree	Elastic Net	Gaussian Process	Gradient Boosting	Hist Gradient Boosting	Huber	KNN	Lasso	Least Absolute Deviations
Ischemic Stroke 30-D..	mean	0.3479	0.2484	0.2924	0.2538	0.2714	0.3381	0.2528	0.2537	0.2758	0.2538	0.2538
	std	0.0304	0.0248	0.0419	0.0261	0.0249	0.0348	0.0258	0.0259	0.0212	0.0261	0.0261
Stockport Local Heal..	mean	9.9199	7.7178	12.3002	11.7245	13.7244	13.8377	11.3006	8.1307	11.1776	11.7245	11.7245
	std	1.9026	1.1205	2.2317	2.5726	3.8125	2.6939	2.4427	1.0713	2.6793	2.5726	2.5726
All Payer In-Hospita..	mean	3.0739	0.6891	3.4372	4.2331	0.7931	3.7622	3.2543	0.6783	2.0344	4.2331	4.2331
	std	0.7053	0.3139	0.7598	0.7843	0.7339	0.8971	0.7223	0.2927	0.4079	0.7843	0.7843
Mortality from Stroke	mean	30.7010	14.6256	168.3347	151.9811	22.8768	184.8604	157.8416	13.7722	189.1336	153.8115	153.8115
	std	3.9797	6.5186	29.8448	30.6033	13.9470	50.6242	56.8231	7.5527	50.9774	30.8195	30.8195

Dataset		Linear	Multi-layer Perceptron	Ordinal	Orthogonal Matching Pursuit	Passive Aggressive	Random Forest	Ridge	SGD	Support Vector	TheilSen	XGBoost
Ischemic Stroke 30-D..	mean	92837077518	0.2534	0.3419	0.2367	0.2981	0.2532	0.2445	0.2670	23292.7133	2150614128	0.3059
	std	84643282853	0.0260	0.0347	0.0230	0.0369	0.0259	0.0236	0.0230	1151.4928	740869974	0.0295
Stockport Local Heal..	mean	35.4377	33.2126	7.9362	8.4499	13.0514	10.4977	7.9064	33.2407	2492.6936	7.9629	12.6588
	std	2.8588	3.1684	1.1034	1.3876	2.9104	2.2494	1.0735	3.7033	366.4972	1.0881	2.0946
All Payer In-Hospita..	mean	0.7930	13.4008	0.8127	0.6906	0.7433	4.0441	0.7227	14.3248	1405.2665	0.7308	4.4412
	std	0.2223	1.3985	0.3228	0.5168	0.2525	0.7519	0.3441	1.5162	261.6704	0.2646	0.7931
Mortality from Stroke	mean	14.6806	169.6659	43.1387	15.3343	22.5055	141.6118	43.1293	172.4691	1472.2756	14.5775	156.0715
	std	6.4786	41.6808	15.8167	7.7985	3.0215	29.2026	15.8528	42.0313	357.3347	6.6241	38.7280

Table 3: Mean and standard deviation of RMSE for each model across regression datasets.

Huber Regression and Bayesian Ridge Regression emerged as the top performers, achieving the lowest RMSE values. In contrast, SGD Regression exhibited the weakest performance, with the highest RMSE score. Unlike the classification tasks, where the models showed more uniformity, the regression models were more dispersed in their performance – Figure 4 presents a violin plot of the RMSE scores evaluated on the test data from the **Ischemic Stroke 30-Day Mortality and 30-Day Readmission Rates** [5], providing a visual representation of the distribution and variability in model performance. There is a greater variation between models and metrics, with less correlation between them (as shown in Figure 5). This indicates that certain models may perform significantly better than others depending on the data and the evaluation metric used.

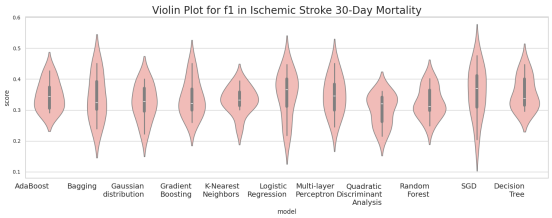

Figure 2: violin plot of the F1 scores from Brain Stroke Dataset

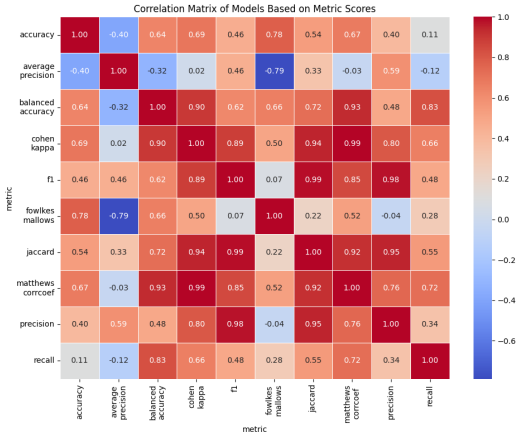

Figure 3: Correlation matrix between classification metrics

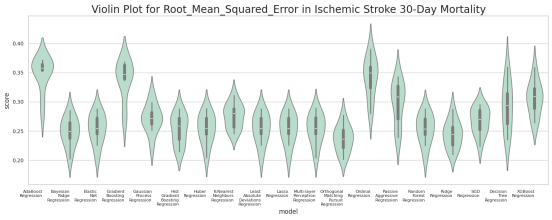

Figure 4: violin plot of the RMSE scores from Ischemic Stroke 30-Day Mortality

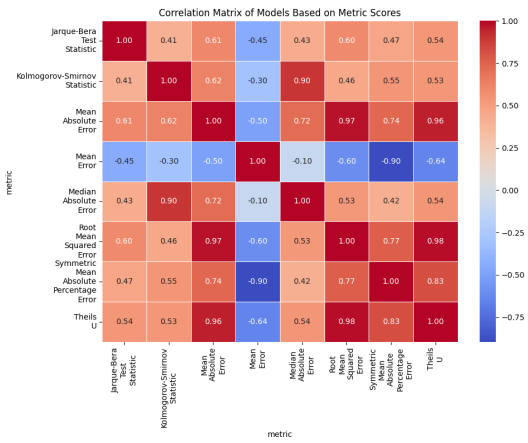

Figure 5: Correlation matrix between regression metrics

5 CONCLUSIONS

In conclusion, our study demonstrates that the performance of AI models in predicting brain stroke outcomes is highly dependent on the quality and characteristics of the datasets used, rather than the choice of the model itself. Through the evaluation of multiple classification and regression models, we observed that while ensemble methods like AdaBoost and Gradient Boosting tended to perform slightly better in classification tasks, the variability between models was minimal across most metrics. However, in the regression tasks, there was a more significant performance dispersion among the models, with some, like Huber Regression and Bayesian Ridge Regression, outperforming others, such as SGD Regression. This suggests that for brain stroke prediction, focusing on the selection of high-quality datasets is essential to enhance model accuracy and reliability.

Furthermore, the study highlights the importance of transparency, reproducibility, and traceability in AI model development for brain stroke analysis. By documenting experimental procedures and datasets in a structured, reproducible format, we can ensure that future research in this area can be independently validated and applied across different patient populations. Our findings emphasize the need for trustworthy, well-curated datasets and standardized methodologies to ensure that AI models in stroke prediction can achieve real-world clinical impact, ultimately improving public health strategies aimed at stroke prevention and recovery.

REFERENCES

[1] Giorgio Colangelo et al. 2024. Prerisk: a personalized, artificial intelligence–based and statistically–based stroke recurrence predictor for recurrent stroke. *Stroke*, 55, 5, 1200–1209.

[2] data.world's Admin. 2021. Stockport local health characteristics. (2021). https://data.world/datagov-uk/0cb6045e-f44f-4dcb-814b-b97840cc80c3?fbclid=IwAR3MEd33szJsu-Sv3aDVuByvmwaBBQhQQw4WYgQG1swlApnYxYKUJBYD7ck.

[3] NHS England. 2022. Mortality from stroke: crude death rate, by age group, 3-year average, mfp. (2022). https://digital.nhs.uk/data-and-information/publications/statistical/compendium-mortality/current/mortality-from-stroke/mortality-from-stroke-crude-death-rate-by-age-group-3-year-average-mfp.

[4] Valery L. Feigin et al. 2023. Pragmatic solutions to reduce the global burden of stroke: a world stroke organization–lancet neurology commission. *The Lancet Neurology*, 22, 12, 1160–1206.

[5] California Health and Human Services. 2018. Ischemic stroke 30-day mortality and 30-day readmission rates and quality ratings for ca hospitals. (2018). https://data.world/chhs/06ed38d3-b047-4ae2-aa00-2e43b5491d6e.

[6] health.data.ny.gov. 2019. All payer in-hospital/30-day acute stroke mortality rates by hospital (sparcs): beginning 2013. (2019). https://data.world/healthdatany/r29i-yr49?fbclid=IwAR03liwBhR_XWfdnkj3tWBKjdHDljDTiY9YiSDSsTXdgwVR7OOxfBQuPNa0.

[7] Tianyu Liu, Wenhui Fan, and Cheng Wu. 2019. Data for a hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical-datasets. Mendeley Data, V1. (2019). DOI: 10.17632/x8ygrw87jw.1.

[8] Sajid Md. 2022. Brain stroke dataset. (2022). https://data.world/researchersj/brain-stroke-dataset?fbclid=IwAR3Y-rrWMYck5OoP15HJVJiihvZVvzVyUj8B7cijBO-Q3XbmQX8fMAd-n0o.

[9] Muhammad Salman Pathan, Zhang Jianbiao, Deepu John, Avishek Nag, and Soumyabrata Dev. 2020. Identifying stroke indicators using rough sets. *IEEE Access*, 8, 210318–210327.

[10] F. Pedregosa et al. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[11] Michele Romoli and Pietro Caliandro. 2024. Artificial intelligence, machine learning, and reproducibility in stroke research. *European Stroke Journal*, 9, 3, 518–520.

[12] Yauhen Statsenko, Fatmah Al Zahmi, Miklos Szolics, and Jamal Al Koteesh. 2022. Prognostication of recovery from acute stroke (pras dataset). (2022). https://data.mendeley.com/datasets/y86srgks26/1.

[13] Wenjuan Wang et al. 2020. A systematic review of machine learning models for predicting outcomes of stroke with structured data. *PLOS ONE*, 15, 6, (June 2020), 1–16.

[14] Yulu Zheng et al. 2022. Rapid triage for ischemic stroke: a machine learning-driven approach in the context of predictive, preventive and personalised medicine. *EPMA Journal*, 13, 2, 285–298.

A CLASSIFICATION VIOLIN PLOTS

This section presents violin plots showcasing various metrics for different models applied to the classification task on the "Brain Stroke Dataset".

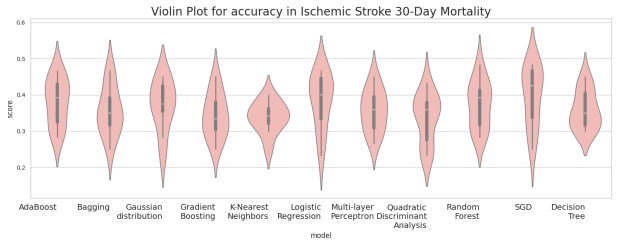

Figure 6: Violin plot of the Accuracy scores from Brain Stroke Dataset

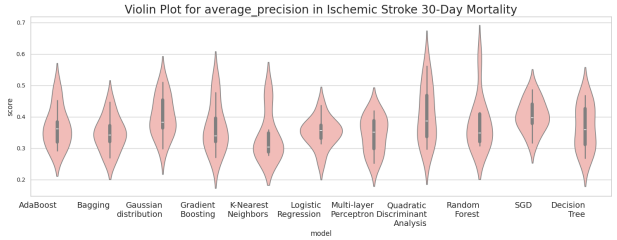

Figure 7: Violin plot of the Average Precision scores from Brain Stroke Dataset

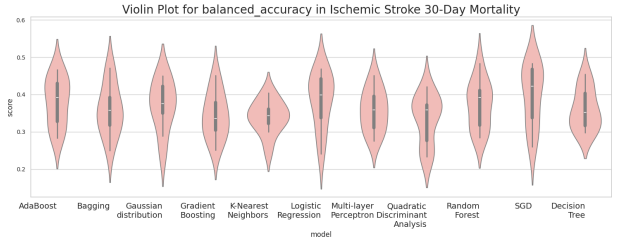

Figure 8: Violin plot of the Balanced Accuracy scores from Brain Stroke Dataset

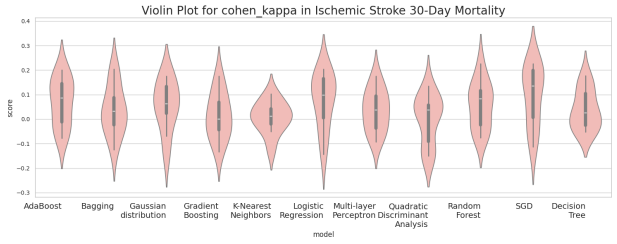

Figure 9: Violin plot of the Cohen's Kappa scores from Brain Stroke Dataset

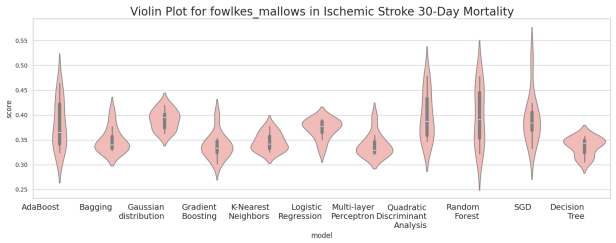

Figure 10: Violin plot of the Fowlkes-Mallows scores from Brain Stroke Dataset

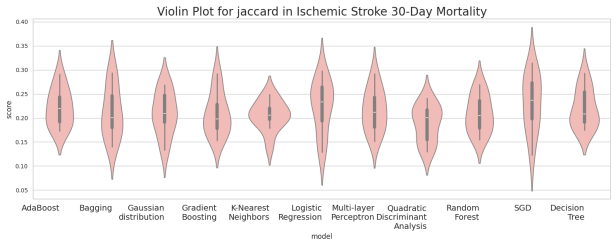

Figure 11: Violin plot of the Jaccard scores from Brain Stroke Dataset

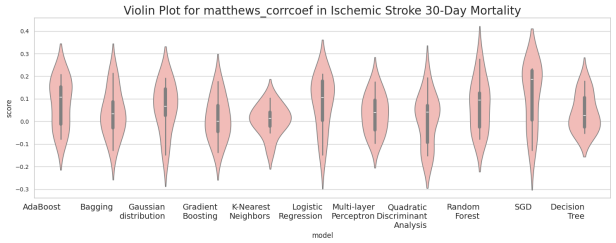

Figure 12: Violin plot of the Matthews Correlation Coefficient (MCC) scores from Brain Stroke Dataset

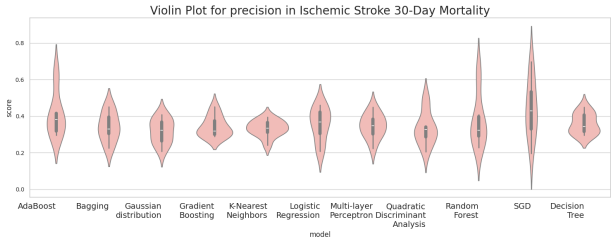

Figure 13: Violin plot of the Precision scores from Brain Stroke Dataset

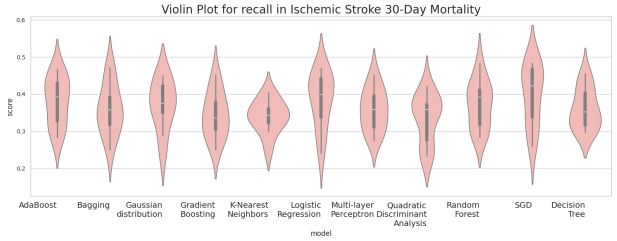

Figure 14: Violin plot of the Recall scores from Brain Stroke Dataset

B REGRESSION VIOLIN PLOTS

This section contains violin plots illustrating various metrics for different models applied to the regression tasks of predicting Ischemic Stroke 30-Day Mortality and 30-Day Readmission Rates.

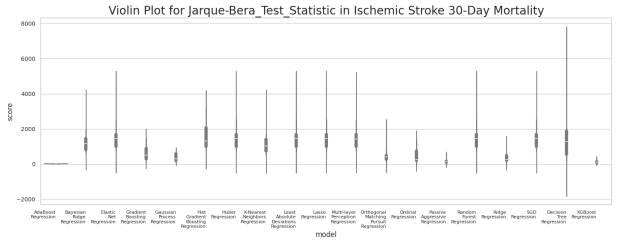

Figure 15: Violin plot of the Jarque-Bera Test Statistic from Ischemic Stroke 30-Day Mortality

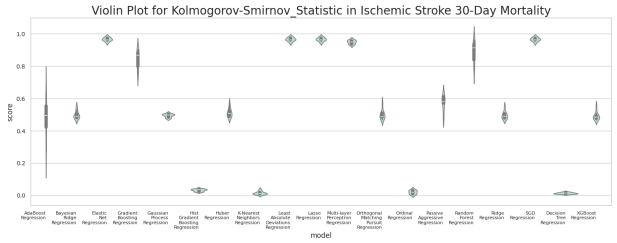

Figure 16: Violin plot of the Kolmogorov-Smirnov Statistic from Ischemic Stroke 30-Day Mortality

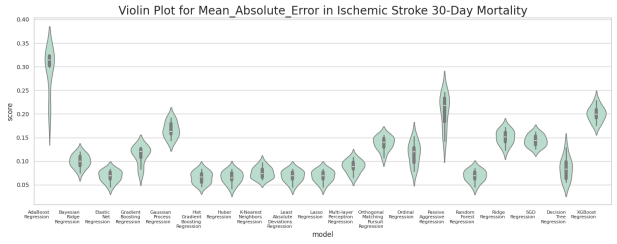

Figure 17: Violin plot of the Mean Absolute Error (MAE) scores from Ischemic Stroke 30-Day Mortality

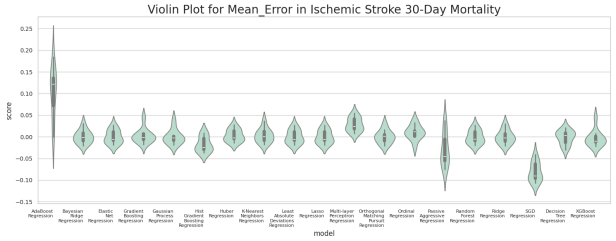

Figure 18: Violin plot of the Mean Error (ME) scores from Ischemic Stroke 30-Day Mortality

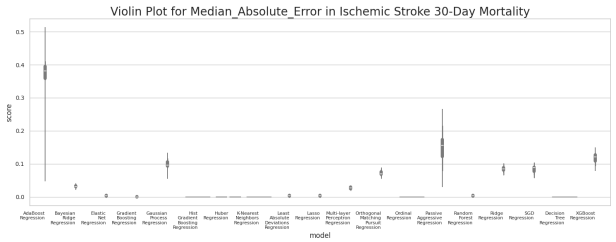

Figure 19: Violin plot of the Median Absolute Error (MedAE) scores from Ischemic Stroke 30-Day Mortality

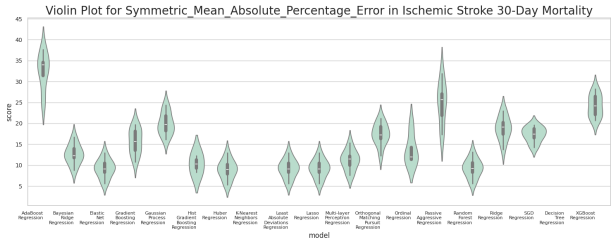

Figure 20: Violin plot of the Symmetric Mean Absolute Percentage Error (SMAPE) scores from Ischemic Stroke 30-Day Mortality

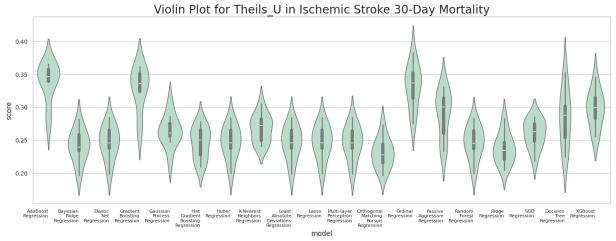

Figure 21: Violin plot of Theil's U from Ischemic Stroke 30-Day Mortality