Stroke Prediction Based on Machine Learning Models

Tianyue Huang, Ziming Song, Hongdao Meng
th3113, zs2815, hm3424

00.00.00

PART 01

Introduction

- Stroke is one of the leading causes of death globally, responsible for approximately 11% of total deaths. Early identification of individuals at high risk is essential to enable preventive measures.
- Our project aims to use Machine Learning methods that we learned in class and develop predictive models that accurately forecast stroke risk based on key clinical and demographic factors, including age, gender, and health conditions.
- Using a variety of machine learning models, from traditional techniques such as **logistic regression**, **support vector machines**, **decision trees**, and **Naive Bayes** to advanced architectures like **Transformers**, we explore different strategies for optimizing stroke prediction accuracy.
- Our work evaluates the models across several metrics, including **recall**, **average precision**, **F1 score**, **accuracy**, **macro averages**, and **weighted averages**, and highlights the potential of hybrid and deep learning models for risk assessment in healthcare.

PART 02

Related Work

Traditional Machine Learning Approaches

1) logistic regression has been a popular baseline method for stroke prediction. However, its linear nature limits its ability to capture complex interactions between medical features.

2)researchers applied random forest models to predict strokes using patient health data, achieving an accuracy of 98.98%

3)SVM has also been explored for stroke prediction comparing different kernel methods to capture non-linear relationships. SVM is sensitive to hyperparameter selection and probably struggles with high-dimensional datasets or class imbalances

Deep Learning for Stroke Prediction

1) CNN leveraging MRI images successfully predicted ischemic stroke lesion size and location with an AUC of 0.92.

2)RNNs have been applied on sequential health data, such as patient vitals and medication history, improving temporal pattern recognition in stroke prediction tasks

3)Transformer-based architectures, initially designed for natural language processing, are increasingly being applied in medical area due to their good performance on modeling long-range dependencies.

Hybrid Models

Researchers explored integrating CNN features extracted from image data with random forests models to predict stroke risk, achieving superior performance compared to standalone models

PART 03

Dataset

Stroke Prediction Dataset with 11 clinical features for predicting stroke events (https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset)

Attribute name	Details
id	unique identifier
gender	"Male", "Female" or "Other"
age	age of the patient
hypertension	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
heart_disease	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
ever_married	"No" or "Yes"
work_type	"children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
residence_type	"Rural" or "Urban"
avg_glucose_level	average glucose level in blood
bmi	body mass index
smoking_status	"formerly smoked", "never smoked", "smokes" or "Unknown"
stroke	1 if the patient had a stroke or 0 if not

PART 04

Accuracy/Error Measures

Recall: Recall measures the proportion of relevant labels that are correctly identified by the model

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

Average Precision (AP): A higher AP indicates that the model can accurately recognize relevant labels in a multi-label context. Specifically, AP is helpful in understanding how well the model identifies positive instances when the data is imbalanced.

$$\text{AP} = \sum_n (R_n - R_{n-1}) \cdot P_n$$

F1 Score: The F1 score is the harmonic mean of precision and recall, which provides a balance between the two metrics.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Accuracy: Accuracy measures the proportion of correctly predicted instances among all instances in the dataset.

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{Total Instances}}$$

Macro Average: Macro average computes the unweighted mean of precision, recall, and F1 scores across all classes. This metric treats all classes equally, regardless of their support (i.e., class size).

$$\text{Macro F1} = \frac{1}{C}\sum_{i=1}^{C}\text{F1}_i$$

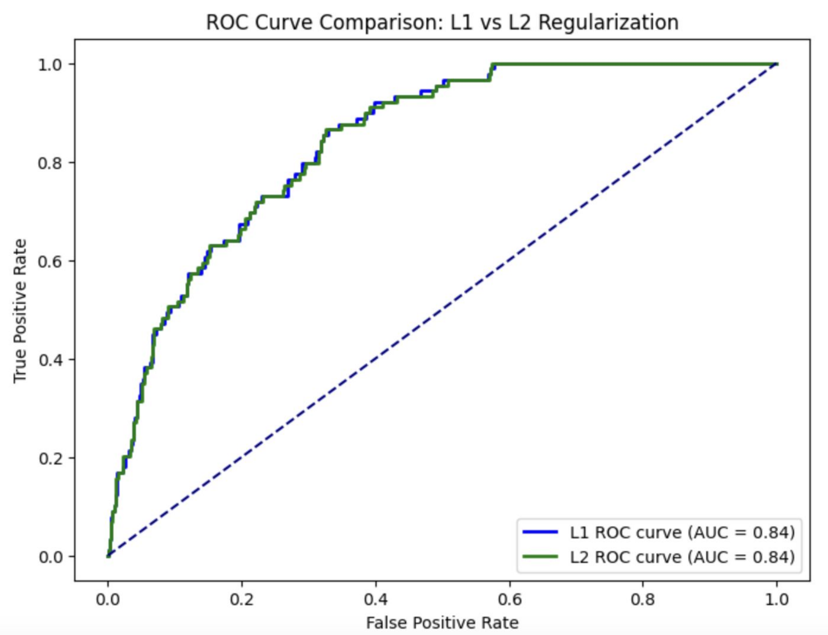
Weighted Average: Weighted average calculates the mean of precision, recall, and F1 scores across all classes, weighted by the number of true instances in each class.

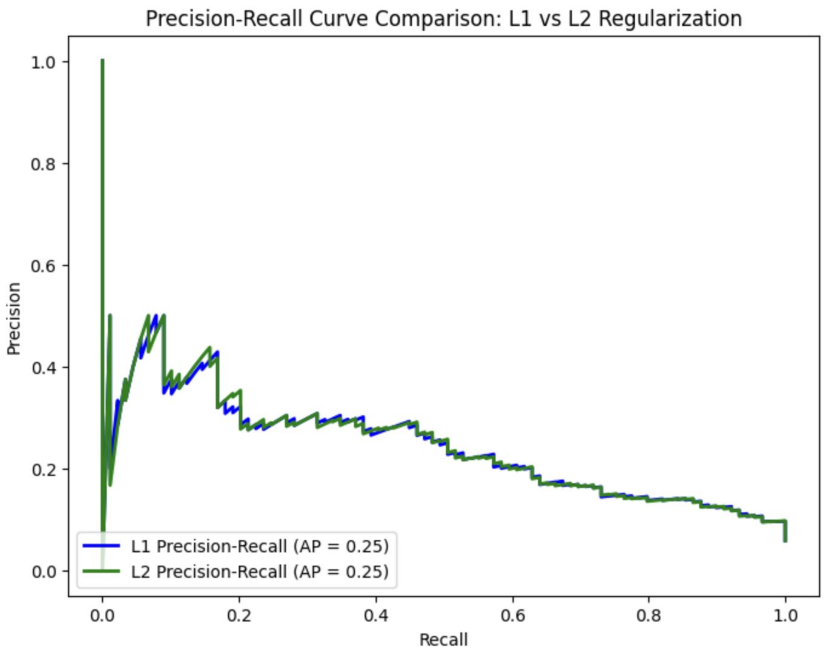
$$\text{Weighted F1} = \sum_{i=1}^{C} w_i \cdot \text{F1}_i$$

PART 05

ML Models and Analysis

logistic regression (baseline)





The models struggle to correctly predict the minority class.

SVM

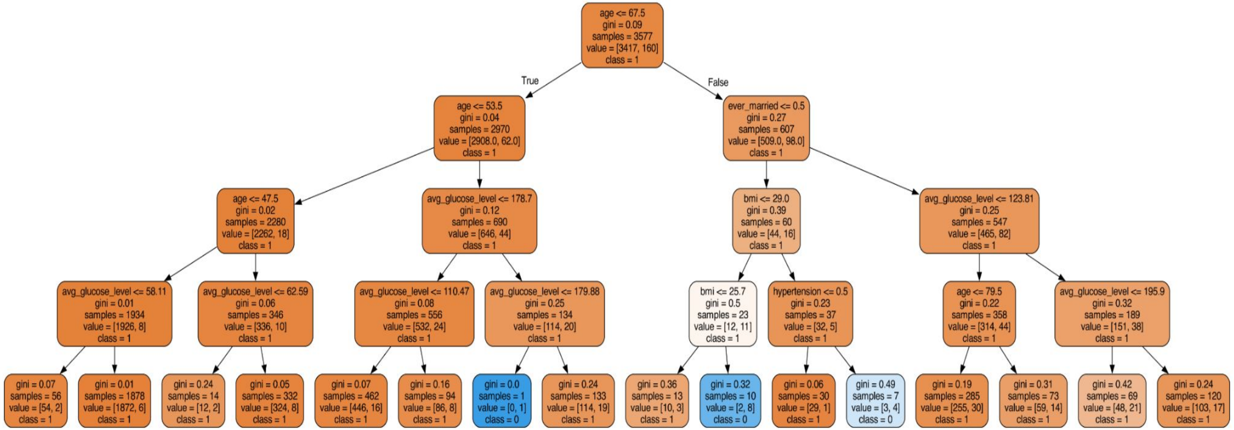
The SVM model is evaluated using four different kernels respectively: linear, polynomial (poly), radial basis function(rbf), and sigmoid. Among these, the poly and rbf kernels demonstrated the best overall performance, especially on the minority class (Class 1)

		Precision	Recall	F1-Score	Support
0	Linear	0.76	0.67	0.71	42
	Ploy	0.76	0.69	0.72	42
	Rbf	0.79	0.62	0.69	42
	Sigmoid	0.54	0.60	0.57	42
1	Linear	0.70	0.79	0.74	42
	Ploy	0.72	0.79	0.75	42
	Rbf	0.69	0.83	0.75	42
	Sigmoid	0.55	0.50	0.53	42
Accuracy	Linear			0.73	84
	Ploy			0.74	84
	Rbf			0.73	84
	Sigmoid			0.55	84
Macro avg	Linear	0.73	0.73	0.73	84
	Ploy	0.74	0.74	0.74	84
	Rbf	0.74	0.73	0.72	84
	Sigmoid	0.55	0.55	0.55	84
Weighted avg	Linear	0.73	0.73	0.73	84
	Ploy	0.74	0.74	0.74	84
	Rbf	0.74	0.73	0.72	84
	Sigmoid	0.55	0.55	0.55	84

Decision tree

	Precision	Recall	F1-Score	Support
0	0.95	0.94	0.95	1444
1	0.18	0.20	0.19	89
Accuracy			0.90	1533
Macro avg	0.56	0.57	0.57	1533
Weighted avg	0.91	0.90	0.90	1533

The model achieved an overall accuracy of 89.89%. The model performs well on the majority class, with high precision, recall and f1-score, which it still struggles with in the minority class.



NYU

Naive Bayesian

	Precision	Recall	F1-Score	Support
0	0.86	0.29	0.43	42
1	0.57	0.95	0.71	42
Accuracy			0.62	84
Macro avg	0.71	0.62	0.57	84
Weighted avg	0.71	0.62	0.57	84

The Gaussian Naive Bayes model shows strong recall for class 1 (stroke cases) at **95%**, effectively identifying most stroke cases. However, its precision for class 1 is only **57%**, which indicates many false positives. For class 0 (non-stroke cases), the model achieves high precision (**86%**) but struggles with recall (**29%**), which leads to frequent misclassification of non-stroke cases as strokes. With an overall accuracy of **62%**, the model performs better on the minority class but struggles to balance precision and recall across both classes, as reflected in its low F1-scores for both classes.

Transformer

	Precision	Recall	F1-Score	Support
0	0.98	0.78	0.87	973
1	0.13	0.65	0.21	49
Accuracy			0.77	1022
Macro avg	0.55	0.71	0.54	1022
Weighted avg	0.94	0.77	0.83	1022

The Transformer-based binary classifier achieved a high accuracy of **93.84%**, which indicates strong overall performance. However, it struggled with class 1 (stroke cases), achieving a recall of **65%** but a low precision of **13%**, resulting in a poor F1-score of **0.21**. This suggests that while the model identifies many stroke cases, it also generates a high number of false positives.

PART 06

Overall Results

		logistic regression		SVM		Decision tree	Naive Bayesian	Transformer
Average Precision		L1	0.94	Linear	0.7487	0.9419	0.6056	0.9384
				Ploy	0.7410			
		L2	0.94	Rbf	0.7519			
				Sigmoid	0.6003			
Recall	Macro Avg	L1	0.50	Linear	0.73	0.57	0.62	0.71
				Ploy	0.74			
		L2	0.50	Rbf	0.73			
				Sigmoid	0.55			
	Weighted Avg	L1	0.94	Linear	0.73	0.90	0.62	0.77
				Ploy	0.74			
		L2	0.94	Rbf	0.73			
				Sigmoid	0.55			
F1 Score	Macro Avg	L1	0.49	Linear	0.73	0.57	0.57	0.54
				Ploy	0.74			
		L2	0.49	Rbf	0.72			
				Sigmoid	0.55			
	Weighted Avg	L1	0.91	Linear	0.73	0.90	0.57	0.83
				Ploy	0.74			
		L2	0.91	Rbf	0.72			
				Sigmoid	0.55			

PART 07

Conclusion

1. Model Performance and Class Imbalance

Across all models, class imbalance significantly influenced performance, particularly for identifying stroke cases (the minority class). Logistic regression and SVM achieved competitive accuracy (94.19% and 93.84%, respectively), but both struggled with precision and recall for stroke cases. While Naive Bayesian models excelled in recall for stroke cases (95%), they suffered from low precision (57%), leading to frequent false positives. Similarly, decision trees and Transformers displayed strong performance on non-stroke cases but fell short in reliably identifying stroke cases, reflecting the overarching challenge of handling imbalanced datasets effectively.

2. Transformer Potential

The Transformer-based model demonstrated its ability to handle complex feature interactions with a high overall accuracy (93.84%) and strong performance for non-stroke cases (precision: 98%, F1-score: 0.87). However, it struggled with the minority class, achieving only 13% precision and an F1-score of 0.21 for stroke cases. This phenomenon emphasizes that while advanced models like Transformers can improve overall predictive power, they still need additional refinement—such as balancing techniques or domain-specific adjustments—to effectively handle predictions for minority classes.

3. Future Directions

To improve stroke prediction models, addressing class imbalance is critical. Techniques like oversampling, undersampling, and class-weighted loss functions should be prioritized. Additionally, simpler models like logistic regression remain reliable baselines, but advanced models such as Transformers offer promise if equipped with hyperparameter tuning and tailored feature engineering. Future efforts should focus on optimizing recall and precision for stroke cases, as early and accurate detection of strokes is crucial for implementing timely medical interventions and reducing mortality rates.

PART 08

References

1. Hosseini, M. S., & Luo, J. (2017). *Stroke risk prediction using logistic regression and electronic health records*. Journal of Medical Systems, 41(10), 170.
2. Random Forest for Stroke Prediction (2020): https://jamanetwork.com/journals/jamanetworkopen/article-abstract/2762679
3. Zhang, Y., & Wallace, B. C. (2015). *A sensitivity analysis of SVMs in text classification*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(5), 1100-1111.
4. Deep Learning for Ischemic Stroke Lesion Prediction (2021): https://ieeexplore.ieee.org/document/10284685
5. Lipton, Z. C., Kale, D. C., Elkan, C., & Wetzel, R. (2016). *Learning to diagnose with LSTM recurrent neural networks*. arXiv preprint arXiv:1511.03677.
6. Song, H., Rajan, D., Thiagarajan, J. J., & Spanias, A. (2018). *Attend and diagnose: Clinical time series analysis using attention models*. AAAI.
7. Suk, H. I., Lee, S. W., & Shen, D. (2017). *Deep ensemble learning of sparse regression models for brain disease diagnosis*. Medical Image Analysis, 37, 101-113.
8. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: Synthetic Minority Over-sampling Technique*. Journal of Artificial Intelligence Research, 16, 321-357.
9. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., ... & Chen, T. (2018). *Recent advances in convolutional neural networks*. Pattern Recognition, 77, 354-377.
10. Li, W., Milletarì, F., Xu, D., & et al. (2019). *Privacy-preserving federated brain tumour segmentation*. Medical Image Analysis, 61, 101663.
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
12. Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. arXiv preprint arXiv:1607.06450.

NYU

PART 08

Thank you!

Template Notes

This template is part of the **NYU Templates collection**. Refer to our **Usage Guidelines** for help topics and quick tips on how to use this template.

Download the **Grackle Slides** add-on to automatically run accessibility checks on all aspects of your document and get advice on how to make things better.