Capstone Project

STROKE DIAGNOSIS FROM PATIENT DATA USING MACHINE LEARNING

ADESOLA ADEWALE OMONIYI

8th November, 2023

INTRODUCTION

- Stroke is a medical emergency that occurs when the blood supply to the brain is interrupted or damaged, causing brain cells to die and affecting various functions of the body and mind.
- Challenges of stroke diagnosis
  - Complex risk factor
  - Incomplete/Inaccurate
- Importance of timely and accurate diagnosis
  - Patients: Avoids serious complications
  - Healthcare: reduced cost and resource allocation

LITERATURE REVIEW

- Patil et al. (2022) used a synthetic dataset and an MLP classifier to achieve an accuracy score of 0.88, a precision of 0.89, a recall of 0.87, and an f1-score of 0.88 for stroke diagnosis, but suggested that other metrics and techniques may be needed to improve the model.
- Alegiani et al. (2020) used a large dataset of RCTs and a random forest classifier to achieve an accuracy score of 0.82, a precision of 0.83, a recall of 0.81, and an f1-score of 0.82 for assessing the quality of stroke patient information, but indicated that other methods and metrics may be needed to improve the model.
- Potdar et al. (2021) reviewed 25 studies that used various ML algorithms for stroke disease classification and prediction, and found that the average accuracy score of the ML algorithms for stroke diagnosis was 0.91 and for stroke prognosis was 0.86, and discussed the pros and cons of each ML algorithm.

LITERATURE REVIEW

- Ahmad (2023) worked on a similar dataset and obtained an F1 score of 84% and ROC-AUC score of 83% using the XGB classifier
- He also used the K neighbor classifier to obtain an F1 score of 89% and ROC-AUC score of 98.8% using the XGB classifier
- Comparison of previous studies with our proposed project (stroke prediction | Imbalanced dataset | Kaggle)
- Tasnim (2023) created a model using the artificial neural network to predict the likely hood of having stroke from the same dataset and achieved an accuracy of about 95% without computing the f1 score (StrokePediction_ANN | Kaggle).

PROPOSED APPROACH

- Overview of our data science approach
  - Train 4 different models
  - Fine tune model with the highest f1 score
- Description of the dataset
  - Contains 5,110 observations with 11 features
  - Target has imbalance class of 95:5
- The Random Forest classifier
  - ML algorithm that uses multiple decision trees, each trained on a random subset of the data and features, and combines their predictions by voting or averaging to improve the accuracy and reduce the overfitting of the model
- Choice of F1 score as the evaluation metric
  - Harmonic mean of recall and precision

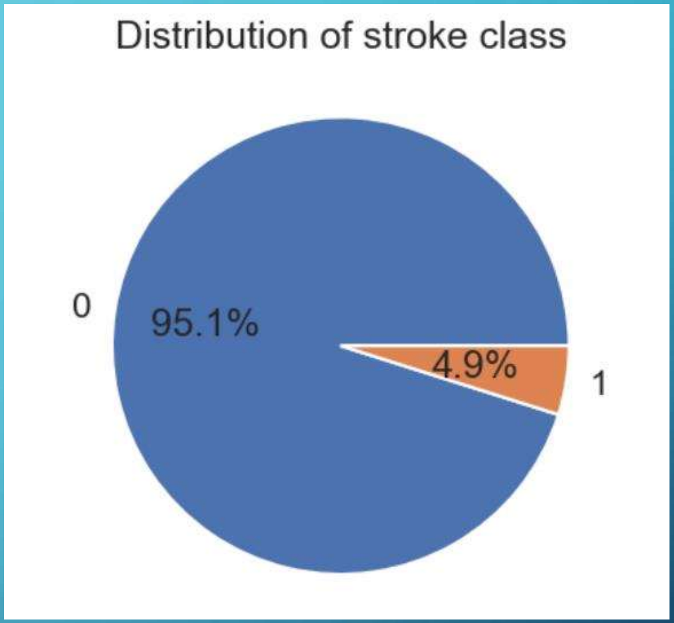

Figure 1: Stroke Distribution

DATA PREPROCESSING

- Handling missing values
  - BMI
- Handling outliers
  - BMI
  - Average glucose level
- Data normalization and feature scaling
  - Standard scaler
- Feature engineering and selection
  - Id
  - Marital status

MODEL TRAINING

- Splitting the data into training and testing sets
- Training
  - Logistic Regression
  - Random Forest
  - Decision Tree
  - K Neigbhour
- Over Sampling
  - Smote
  - Borderline smote
  - Random Oversampler
- Hyperparameter tuning and optimization

EVALUATION

- Assessing the model's performance on the testing set
  - test precision of 92.2%
  - test recall of 96%
  - test f1 score of 94%
  - test roc auc of 98.8%
- Interpreting the F1 score metric
- Analyzing the model's strengths and weaknesses

CONCLUSION

- Summary of key findings and implications
  - the model achieves a good balance between precision and recall, indicating its overall effectiveness
  - Age, bmi and average glucose level are key in predicting stroke
  - As one ages (40 years +), the higher the bmi, the more likely one may be stroke positive
  - on the average, the possibility of having stroke increases as one crosses 40 years of age.
- Discussion of future directions and research opportunities
  - Model should undergo rigorous testing
  - Training on a larger dataset
  - Developing other
- Emphasis on the potential benefits of machine learning for stroke diagnosis
  - Early Stroke Risk Identification
  - Clinical Decision Support
  - Resource Allocation and Prioritization:

THANK YOU

REFERENCES

- **References**

- Alegiani, A. C., Rahn, A. C., Steckelberg, A., Thomalla, G., Heesen, C., & Köpke, S. (2020, November 6). *Quality of Stroke Patient Information Applied in Randomized Controlled Trials—Literature Review.* Frontiers. Retrieved November 7, 2023, from https://www.frontiersin.org/articles/10.3389/fneur.2020.526515/full
- Kaggle. (n.d.). *Stroke Prediction Dataset*. Kaggle. Retrieved November 7, 2023, from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
- Potdar, V., Santhosh, L., & Gowda, Y. R. (2021, September 6). *A Survey on Stroke Disease Classification and Prediction using Machine Learning Algorithms – IJERT*. International Journal of Engineering Research & Technology. Retrieved November 7, 2023, from https://www.ijert.org/a-survey-on-stroke-disease-classification-and-prediction-using-machine-learning-algorithms