# Stroke Prediction Model

CS 613 – Machine Learning

Authors: Danny Li, Tien Nguyen, Emily Wang

June 8th, 2021

# Contents


Introduction

Related/Prior Work

Approach

Datasets

Results

Conclusion

Future Work

Reference

# I. Introduction

**Benefits of *earlier treatment*:**

- Give greater chance of surviving and recovery
- Reduce likelihood of permanent disability
- Lesser need for extensive rehabilitation

**Stroke Risk factors:**

- Lifestyle
- Medical factors
- Other factors: age, race, sex, hormones

**➔ Stroke Prediction Model**

# II. Related/Prior Work

## “Stroke Prediction Using SVM” by R.S Jeena and Sukesh Kumar

- Same data set from International Stroke Trial Database (Physiological Features)
- Implemented Support Vector Machine (SVM) with different kernel functions
- Linear Kernel accuracy of **91%**

**2016**

**2020**

## “Machine Learning Approach to Identify Stroke Within 4.5 Hours” by Hynna Lee et all.

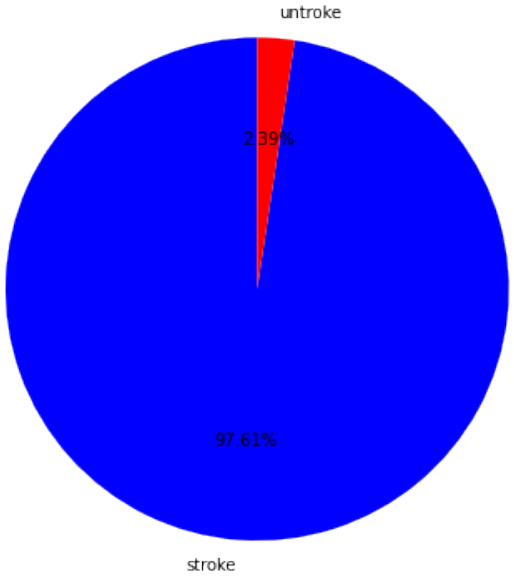
- Analyzed diffusion-weight imaging and fluid attenuated inversion recovery magnetic resonance imaging as data
- Implemented logistic regression, SVM and random forest
- ML algorithms had greater sensitivities (recall) than human readings. (75.8% for logistic regression, 72.7% for SVM, 48.5% for human reading)

# III. Approach

1. Data preprocessing → imbalance datasets and balanced datasets
2. Binary *Logistic Regression, Naïve Bayes,* and *Decision Tree (ID3)* classifiers to train and validate the imbalance datasets.
3. Implement an ensemble technique: *Random Forest* Classifier
4. Binary *Logistic Regression, Naïve Bayes,* and *Decision Tree (ID3)* classifiers to train and validate the balanced datasets.
5. Implement *cross-validation* to improve accuracy and avoid overfitting
6. Analyze and compare statistics results

# III. Datasets

- International Stroke Trial Database(112 features and 19,435 samples)




No-stroke: 2.39%  Stroke: 97.6%

- Data preprocessing
  - 18 features dataset (13,079 samples – imbalanced dataset)
  - 12 paper features dataset(13,079 samples)
  - 12 top features dataset (13,079 samples)
    - Entropy

Recall the formula for entropy:

$$H\big(P(v_1), \ldots, P(v_n)\big) = \sum_{i=1}^{n} -P(v_i) \log_n P(v_i)$$

# Comparison of two datasets

- **12 paper features dataset**
  - ***Sex***
  - **Age**
  - **Symptoms noted on waking**
  - ***Atrial fibrillation***
  - Infarct visible on CT
  - Face deficit
  - Arm/hand deficit
  - Leg/foot deficit
  - ***Dysphasia***
  - Hemianopia
  - ***Visuospatial disorder***
  - Brainstem/cerebellar signs
  - Other deficit

- **12 top features dataset**
  - ***Conscious state at ran***
  - Age
  - Symptoms noted on waking
  - ***CT before randomization***
  - Infarct visible on CT
  - ***Systolic blood pressure at randomization***
  - Face deficit
  - Arm/hand deficit
  - Leg/foot deficit
  - Hemianopia
  - Brainstem/cerebellar signs
  - Other deficit

# V. Results

## Statistics Results on Imbalance Datasets

Table1. Validation Statistics Results for the three datasets

| Model | Dataset | Precision | Recall | f-measure | Accuracy |
|---|---|---|---|---|---|
| Logistic Regression | 18 features | 0.975 | 1 | 0.987 | 0.975 |
| | 12 paper features | 0.975 | 1 | 0.987 | 0.975 |
| | 12 features | 0.975 | 1 | 0.987 | 0.975 |
| Naive Bayes | 18 features | 0.980 | 0.940 | 0.960 | 0.923 |
| | 12 paper features | 0.979 | 0.96 | 0.969 | 0.941 |
| | 12 features | 0.979 | 0.946 | 0.962 | 0.928 |
| Decision Tree (ID3) | 18 features | 0.977 | 0.985 | 0.981 | 0.963 |
| | 12 paper features | 0.975 | 0.999 | 0.987 | 0.973 |
| | 12 features | 0.976 | 0.997 | 0.986 | 0.973 |

**=> Decision Tree and Logistic Regression classifiers give better precision and recall results**

# V. Results

## Random Forest Classifier on Imbalance Datasets

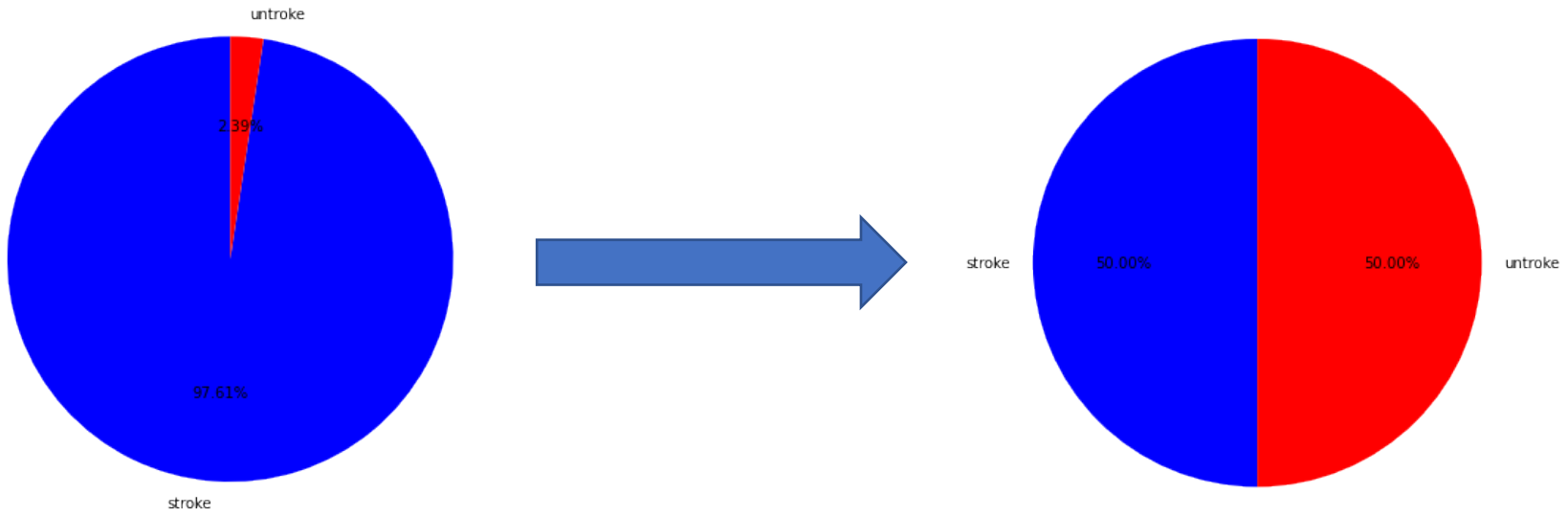
- Dataset: No-stroke: 0.0239, Stroke: 0.976
  => Lowest acceptable accuracy: 0.976
- Previous statistics:
  - Precision: 0.975
  - Recall: 0.999
  - F-measure: 0.987
  - Accuracy: 0.973
- Random Forest:
  - $m = \sqrt{p}$
  - Consistent statistic results for all imbalanced datasets

Table 2. Statistics Results for Random Forest Classifier on 12-paper-features

| Sample Size | Number of Trees | Precision | Recall | f-Measure | Accuracy |
|---|---|---|---|---|---|
| 6000 | 1 | 0.977 | 0.999 | 0.988 | 0.976 |
| | 10 | 0.977 | 1 | 0.988 | 0.977 |
| 8000 | 1 | 0.977 | 0.999 | 0.988 | 0.975 |
| | 10 | 0.977 | 1 | 0.988 | 0.977 |
| | 100 | 0.977 | 1 | 0.988 | 0.977 |
| | 200 | 0.977 | 1 | 0.988 | 0.977 |
| | 300 | 0.977 | 1 | 0.988 | 0.977 |
| | 500 | **0.977** | **1** | **0.988** | **0.977** |

# Balanced Datasets

- Imbalance datasets: 13,079 samples
  - 12,767 stroke
  - 312 no-stroke
- Balanced datasets: 612 samples
  - 312 stroke
  - 312 no-stroke



# V. Results

## Statistics Results on Balanced Dataset (12-paper-features)

| Model | Dataset | Precision | Recall | f-measure | Accuracy |
|---|---|---|---|---|---|
| Logistic Regression | Training | 0.665 | 0.662 | 0.664 | 0.673 |
| | Validation | 0.723 | 0.714 | 0.719 | 0.699 |
| Naive Bayes | Training | 0.722 | 0.528 | 0.610 | 0.68 |
| | Validation | 0.73 | 0.47 | 0.571 | 0.611 |
| Decision Tree (ID3) | Training | 0.716 | 0.906 | 0.8 | 0.815 |
| | Validation | 0.504 | 0.652 | 0.569 | 0.562 |

$\Rightarrow$ Logistic Regression works best on the balanced dataset

$\Rightarrow$ Overfitting with Naïve Bayes and Decision Tree

# V. Results
## Statistics Results of Cross-Validation k-Folds on Balanced Datasets

- Accuracy values are improved for Naïve Bayes and Decision Tree
- Logistic Regression give the highest accuracy

Table 4. Accuracy for Cross-validation of Logistic Regression, Naïve Bayes, and Decision Tree classifiers

| K-folds | Logistic Regression | Naive Bayes | Decision Tree |
|---|---|---|---|
| 5 | 0.513 | 0.569 | 0.516 |
| 10 | 0.59 | 0.607 | 0.555 |
| 50 | 0.662 | 0.657 | 0.578 |
| 100 | 0.669 | 0.665 | 0.585 |
| 500 | 0.682 | 0.674 | 0.594 |

# VI. Conclusion

- Random Forest produced best results with imbalanced dataset
  - Precision of 0.977
  - Recall of 1
- Logistic Regression produced best results with balanced dataset
  - Accuracy of 0.699
  - Precision of 0.723
  - Recall of 0.714
- Better results compared to “Stroke Prediction Using SVM” by Jeena and Kumar (97.7% vs 91%)
  - Possible Causes – same features but different samples after preprocessing dataset, paper does not mention balance or imbalanced dataset

# VII. Future Work

- Explore/Apply other types of machine learning approaches to dataset
  - K-nearest neighbors (KNN)
  - Support vector machine (SVM)
  - Deep learning
- Train different datasets to validate different models

# References


- “What Is CVD?” *World Heart Federation*, 6 May 2021, world-heart-federation.org/what-is cvd/.
- Sandercock, P.A., Niewada, M., Członkowska, A. *et al.* The International Stroke Trial database. *Trials* **12,** 101 (2011). https://doi.org/10.1186/1745-6215-12-101
- T. R. Dawber, G. F. Meadors, and F. E. Moore. Epidemiological approaches to heart disease: The Framingham study. *American Journal of Public Health and the Nation’s Health*, 41:279–286, March 1951.
- P. A. Wolf, R. B. D’Agostino, A. J. Belanger, and W. B. Kannel. Probability of stroke: a risk profile from the Framingham study. *Stroke*, 22:312–318, March 1991
- R. S. Jeena and S. Kumar, "Stroke prediction using SVM," 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2016, pp. 600-602, doi: 10.1109/ICCICCT.2016.7988020.
- Lee, Hyunna, et al. “Machine Learning Approach to Identify Stroke Within 4.5 Hours.” Stroke, vol. 51, no. 3, 28 Jan. 2020, pp. 860–866., doi:10.1161/strokeaha.119.027611. .