1 Evaluating different machine learning models for predicting silent strokes

2 Group Members

Simeon Hornberger

3 Abstract

Strokes cause the fifth largest amount of deaths every year in the US, causing over 160 thousand deaths in the US in 2021 alone. Sometimes strokes are obvious, but some people can also experience silent strokes, where they don't even realize they are having a stroke. Even though those strokes go unnoticed, they can still cause permanent damage to the brain. For this project I will train and evaluate different machine learning models on a dataset that contains samples of different people at different times after they had a stroke. The final goal is to find out if those models can be applied to a dataset of apparently healthy people to find out if they might have had a silent stroke in the past and if so at what time it happened.

4 Introduction

4.1 Problem Motivation

Even though silent strokes go unnoticed, they can still cause permanent damage to the brain. Because of this, it is important to get medical help as soon as possible to limit the amount of damage the stroke is causing to the brain. People who suspect they might have had a silent stroke can get their biological data taken, which serves as input for a machine learning model that predicts if a person has had stroke in the past and if so at what point in time. With this information at hand, they can be treated accordingly by medical professionals.

4.2 Previous work focused on solving this problem

In `https://dl.acm.org/doi/abs/10.1145/1835804.1835830`, Khosla et al. present a machine learning approach for stroke prediction and compare it to the Cox proportional hazards model. They consider feature selection, data imputation and prediction for medical datasets. They introduce an automatic feature selection algorithm based on the conservative mean heuristic. The machine learning methods they investigated outperformed the Cox model for both binary stroke prediction and stroke risk estimation.
Emon et al. (`https://ieeexplore.ieee.org/abstract/document/9297525`) compare the performance of 10 different classifiers to predict strokes for a set of features. To achieve highest accuracy, the results of the base classifiers are aggregated with a weighted voting approach, which leads to an accuracy of 97%.

4.3 Limitations of previous work

While previous approaches trained machine learning models to predict future strokes that have not happened yet, the approach in this paper focuses on training models to predict strokes that have already happened, but they might have went unnoticed.

5 Statement of Contributions

In this paper, we trained and evaluated three different machine learning models: Logistic regression, K-Nearest neighbors and random forest. These models can be applied to data of apparently healthy people to

predict if those people had a stroke in the past and if so how long ago. The code for this project can be accessed under `https://github.com/SimeonHornberger/COMP683_Project`.

6 Methods

6.1 Problem Formulation

The Problem is people not noticing they have experienced a silent stroke. We train different machine learning models on a dataset to predict if and how long ago a person experienced a stroke. After training those models, we can run the models on the dataset of apparently healthy people to identify people that had a silent stroke. It is crucial for those people to find out they had a silent stroke so they can immediately seek medical help to limit the damage caused to their brains.

6.2 Description

The dataset consisted of multiple files, each containing a cell by feature matrix of a person. This person was either healthy or has had a stroke in the past. The files of persons that had a stroke in the past had a timestamp associated with them. These timestamps were 24 hours, 48 hours, 72 hours, 120 hours, 7 days, 14 days, 90 days and 1 year. Together with the healthy label these timestamps made up the labels for the classifiers. Feature selection was done with domain expertise from Professor Stanley and we ended up with 36 features. The data pre-processing was done with trial and error. There were some features which had a lot of zero values, so it was considered to drop those features, but for some classifiers the performance actually improved when keeping those features, so we decided to drop or don't drop on a case by case basis. Furthermore, basically all features had some zero values, so we imputed those features with the mean value for that feature. The data was also scaled. At first the Standard scaler was used, but later the MinMax scaler was used as it brought huge performance gains. For each model we looked at the confusion matrix and classification report. We put a special focus on recall, as it is less damaging falsely predicting a person has had a stroke even though they didn't have one than predicting a person didn't have a stroke even though the person has had one.

6.3 Schematic illustration

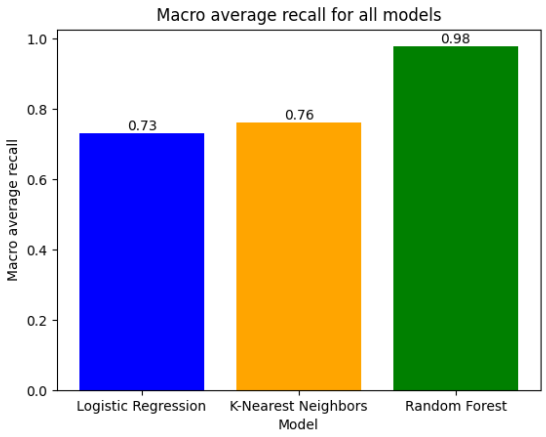
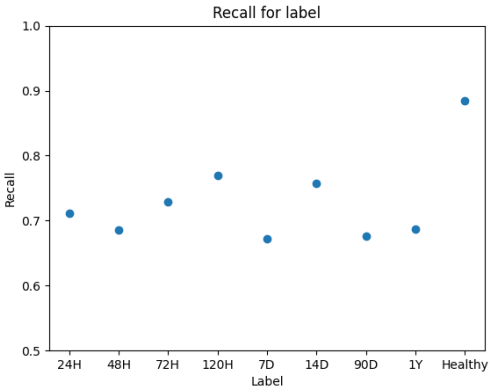


Figure 1: Comparison of macro average for all three models

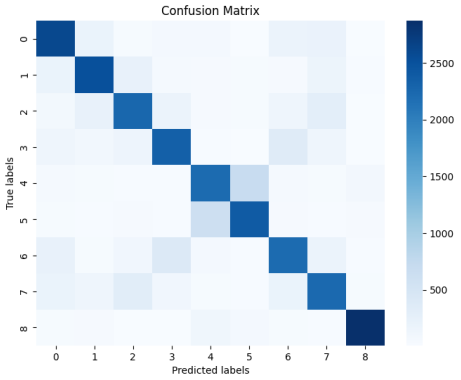
7 Results

7.1 Models

7.1.1 Logistic Regression



(a) Recall for all labels

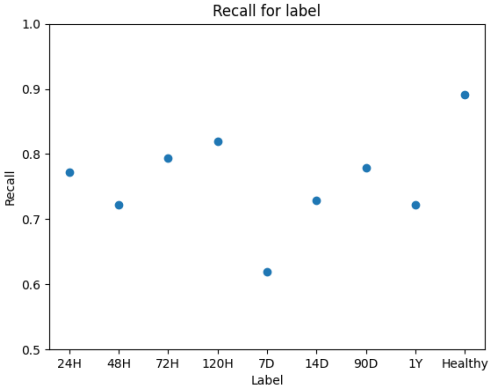


(b) Confusion matrix

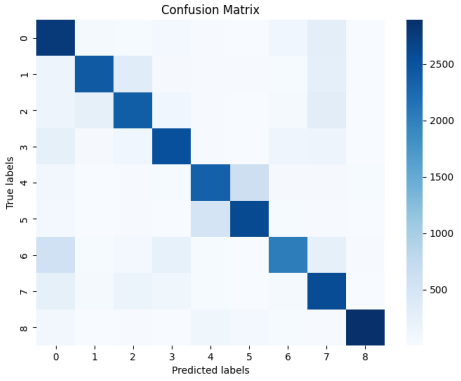
Figure 2: Logistic Regression

For logistic regression, the features with a lot of zero values were left in, as removing them led to worse performance. Imputing the zero values with the mean value for that feature improved recall, precision and f1-score by about 4%. Tuning hyperparameters like the solver, penalty or C value didn't have much impact on the model. The micro and weighted average for precision, recall and f1-score was all 73%. As you can see in Figure 2a, recall fluctuates around 72%, but it is definitely noticeable that the recall for the healthy label is much higher at 88%.

7.1.2 K-Nearest Neighbors



(a) Recall for all labels



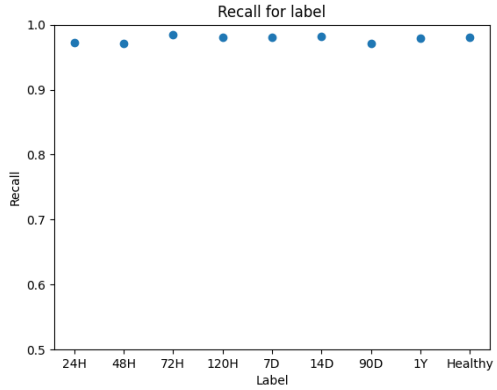
(b) Confusion matrix

Figure 3: K-Nearest neighbors

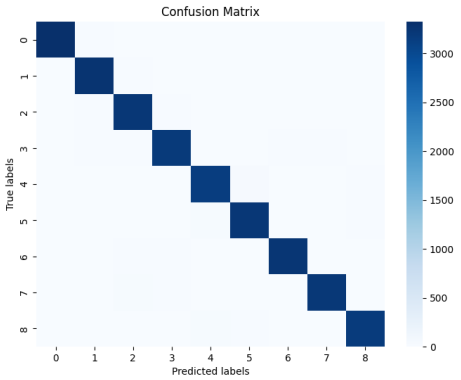
By changing the neighbor hyperparameter for K-Nearest neighbors from 5 to 30 there was a slight performance gain. A much bigger difference made the change from scaling the data with the StandardScaler to

scaling the data with the MinMaxScaler, which led to a performance gain of about 20% for recall, precision and f1-score. Recall fluctuates around 75%, and the recall for the healthy label at 89%. Precision for this model is at 77% and recall and f1-score are at 76%, which is a slight increase compared to logistic regression.

7.1.3 Random Forest



(a) Recall for all labels



(b) Confusion matrix

Figure 4: Random Forest

The random forest model seemed to be the most robust to scaling, as the performance for standard scaling and MinMax scaling were pretty similar, the StandardScaler even being slightly better. A huge performance gain of over 25% could be achieved by changing the hyperparameter of maximal depth from 2 to 20. Another 20% improvement could be obtained by imputing the zero values of a feature with the mean value for that feature. The model also performed slightly better when the features with a lot of zero values were dropped. Overall, the random forest model performed by far the best with a steady recall of 97-98% and average precision and f1-scores of 98%.

7.2 Datasets

A mass cytometry longitudinal dataset about longitudinal recovery of patients from stroke from FlowRepository was used to train and validate the machine learning models(`https://flowrepository.org/id/FR-FCMZYSB`).

7.3 Baselines

There is no real baseline to which the performance of the three models in this paper is compared to, instead we tried to maximize the performance of each model and compared them with the other models.

7.4 Description of Experiments

To evaluate the performance of each classifier, we looked at the classification report and confusion matrix. The classification report includes precision, recall and f1-score. We put a strong emphasis on recall as it is less damaging to falsely predict a person has had a stroke even though the person didn't have one than predicting a person didn't have a stroke when the person has had one.

8 Discussion

8.1 Recap

In this paper we evaluated and compared three different machine learning models to predict if a person has had a stroke in the past or not and if so at which point in time. The used classifiers were logistic regression, K-Nearest neighbors and random forest.

8.2 Observations

The random forest model performed by far the best with an average precision, recall and f1-score of 98%. K-Nearest neighbors and logistic regression were more similar in performance, with K-nearest neighbors slightly in the lead by about 3%. Logistic regression achieved metrics of 73% and K-nearest neighbors 76-77%. For such a delicate matter as predicting silent strokes, it is probably for the best to stick with the random forest classifier and not go with KNN or logistic regression, as you want to be sure that your classifier is right most of the time.

8.3 Limitations and Future Work

What hasn't been considered in this paper are silent strokes that have occured significantly longer ago than 1 year, for example a couple of years ago. Would the classifier label those cases as 1Y or healthy or something completely different. So it would definitely make sense to look into data from strokes where the stroke occured over one year ago and see how the classifiers perform. Another interesting direction would be to predict the type of stroke, whether it is an ischemic or hemorrhagic stroke. But in order to investigate this we would need an extended dataset that also includes the type of stroke for every person. This could be very helpful for people who have experienced silent strokes as they could immediately get the right medical treatment.

8.4 Inspiring Concluding Paragraph

The research that went into this paper showed us that a random forest classifier is by far the best model to predict silent strokes. If people who suspect they might have had a silent stroke run their data through the provided model, they could get the medical help they need and limit the damage to their brain.