**Machine Learning Report (4 pages)**

**Abstract 60words**

Most strokes are caused by an unforeseen obstruction of courses by both the brain and the heart. Early detection and proper management of possible warning signs of stroke may help to minimize further damage. In this paper, machine learning algorithms are applied to various parameters that are used as risk factors to build a model for the prediction of stroke.

**Introduction 212 words**

The World Health Organisation (WHO) claims that stroke is the second leading cause of death worldwide, accounting for approximately 11% of total deaths (WHO, 2020). Stroke is a health condition caused by an insufficient supply of blood (due to lack of oxygen and nutrients) to the brain cells, that further damages them and causes adult disabilities or death. Blood circulation may be disrupted by a clot in a blood vessel or a blood vessel rupture (Jeena). To avoid more damage to the affected region of the brain and complications in other body parts, early diagnosis and careful management are necessary. Predictive machine learning models could be beneficial for patients as they may indicate awareness towards future risks of stoke. Therefore, with the use of machine learning and prediction of outcomes of whether a patient is likely to suffer from stroke, lives of many individuals may be saved with early diagnosis.

The aim of this paper is to find out whether the chosen model and baselines can successfully predict the chances of a patient getting a stroke using the Random Forest algorithm as the main model, along with Support Vector Machine (SVM), K-Nearest Neighbour as baselines to evaluate and compare if the baselines are as good as or rather outperform the main model.

**Literature Review 569 words**

Patients commonly suffer from strokes due to poor lifestyle choices. Consequently, prediction of risks of future strokes for patients has been in high demands as they can be essential for timely prevention and treatments. Generally, in studies and research papers, risk models of stroke have been proposed; however, they typically focus on areas such as patients in specific groups (for example, patients within specific age range or specific locations). Zheng et al. (2015) aimed to develop a predictive model for risks of stroke in future one year’s time period for patients across all age, all disease groups and all payor in Maine. They built a logistic regression model for risk prediction based on multivariate analysis with c-statistics for retrospective and prospective predictions being 0.892 and 0.887 respectively, at a positive predictive value (PPV) of 0.262. Hence, indicating that the model correctly identified 41.0% of risks of stroke for patients in future one year. Similarly, Jeena R S investigate the several physiological parameters used as risk factors for stroke prediction.  Their data was obtained from the International Stroke Trial database and effectively trained and tested by using Support Vector Machine (SVM). Their research shows the implementation of SVM with various kernel functions and concluded that the linear kernel provided an accuracy of 90%.

In another paper, in Chun-Cheng’s article, the physiological data of patients was used to construct an intelligent model that can predict stroke with artificial neural networks (ANNs). They divided the data into training and testing sets, with 43400 and 18601 records, respectively; out of which only training dataset was used. And before the model was built, the dataset was randomly divided into three subsets, which were: 70% training set, 15% test set, and 15% validation set data. When the data was not pre-processed, the accuracy was 98% in 1 hidden node, 50 hidden nodes as well as different algorithms. Therefore, the proposed ANN method in their paper could reach around 98% of accuracy in classification, under the condition of 1000 times cross-validation.

On the other hand, Gracia's study compares to show how different machine learning based modelling techniques, when combined with non-invasive monitoring technologies, may be able to assist in the diagnosis of the category of stroke just a few minutes after the onset of the crisis. This technique has also been tested to predict potential complications, such as a patient's eventual death. Seven different algorithms including: decision tree, K-NN, logistic regression, naive bayes, neural network, random forest, and support vector machines were used and tested using six different metrics. The performances of each model were evaluated using 10-fold cross validation resampling method. In comparison to the other algorithms, random forest models performed the best in diagnosis of stroke and death prediction, with average values of 0.93±0.03 and 0.97±0.01, respectively. Likewise, Emon proposed a study for stroke prediction where different machine learning approaches were used with occurrence of hypertension, heart disease, body mass index level, age, previous stroke, average glucose level and smoking status. The classifiers trained against these high feature attributes included: K-Neighbours Classifier, Gradient Boosting Classifier, Gaussian Classifier, Quadratic Discriminant Analysis, Multi-layer Perceptron Classifier, Logistics Regression, Stochastic Gradient Descent, Decision Tree Classifier, AdaBoost Classifier, and XGBoost Classifier. Thereafter, the results of the base classifiers were combined using the weighted voting method to achieve the highest level of accuracy. Furthermore, their study had a 97% accuracy rate with the weighted voting classifier outperforming the base classifiers.

**Methodology 274 words**

1. Model

Random forest algorithm is an ensemble learning method (also known as ensemble of decision trees), which is commonly used in regression and classification models. Random forests perform by assembling a multitude of decision trees during the time of training process. When it is used in classification, the outcome is the class that is the mode of the individual trees' classes (gracias). The objective of random forest algorithm is to resolve the overfitting decision trees to their training set.

1. Baselines

Based on a set of input samples and corresponding class labels, the purpose of pattern classification is to limit the implicit relation among patterns of the same classification, such that when a test sample is given, the corresponding output class label is retrieved (Jeena). Support vector machine (SVM) is a helpful model for multi-classification problems. It transforms the data using a technique known as the kernel trick, then finds an optimal boundary between the possible outputs based on these transformations, hence making it a dominant tool in data mining and medical image applications.

On the other hand, K-nearest neighbour (KNN) is a non-parametric method that assumes that objects which are identical exist close together, meaning that that similar objects are classified and kept close together. In the feature space, the input consists of the K closest training instances. The output is dependent on KNN's use for regression or classification. This paper utilises KNN classification, which means the output is a class membership. By a majority vote of its neighbours, an instance is classified along with the instance that is assigned to the K-nearest neighbours. Also, K is a small positive integer (Gracias).

**Experiments 739 words**

1. Dataset

The dataset for this works is acquired from a platform for open-source datasets- Kaggle (Kaggle). The dataset collected is available for use to predict whether there is the possibility that a patient will get a stroke based on 5110 data instances as well as the 12 input parameters including: id, gender, age, hypertension, heart disease status, ever married, work type, residence type, average glucose level, BMI, smoking status, and status of ever having stroke, respectively. The dataset was a blend of numerical and categorical data, therefore, being good enough for analysis and prediction generation. Moreover, the dataset had one column- “bmi” which had 201 missing values, and were filled in the data pre-processing stage with the mean value of the column “bmi” using pandas’ “fillna” method. The categorical data were converted to numerical values using the “get\_dummies” feature (also from pandas). Moreover, the dataset was balanced using over-sampling method.

1. Experimental settings

The dataset was split by 70/30 ratio, that is 70% train data and 30% test data. The model and both baselines were applied to the dataset, then the accuracy score and cross validation score were checked to find out the initial responses from the performance of each algorithm for test and train data. The model and baselines were optimised with hyperparameter tuning using grid search technique which attempts to compute the most optimum values of hyperparameters. The hyperparameters tuned for random forest consisted of maximum depth of tree ranging from 2 to 16 and the experimental values for the minimum number of samples required to be at a leaf were ranging from 2 to 10. The hyperparameters were set in SVM by giving the range of 0.1, 1, 10, 100, 1000 to the value of C (controls errors). And for KNN - the number of neighbors to inspect were set to range from 1 to 25.

1. Evaluation criteria

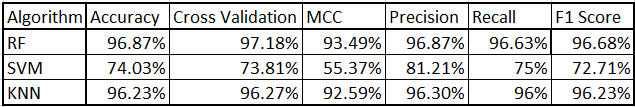
To quantify performances of the model and the baselines, several evaluation techniques to compare the set of algorithms by measuring the confusion matrix, accuracy, cross validation, Matthew’s correlation coefficient, precision, recall (sensitivity) and F1 score were utilised during the experiments. Following are the defined metrics:

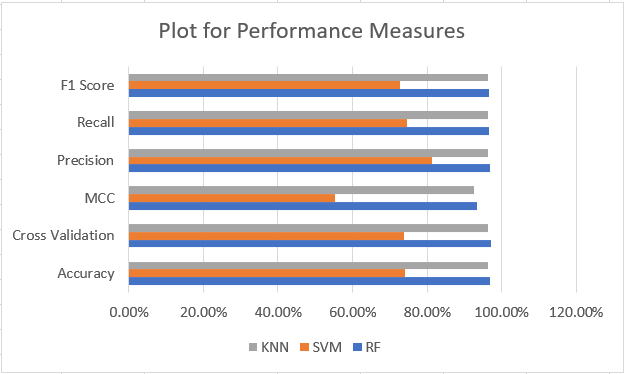
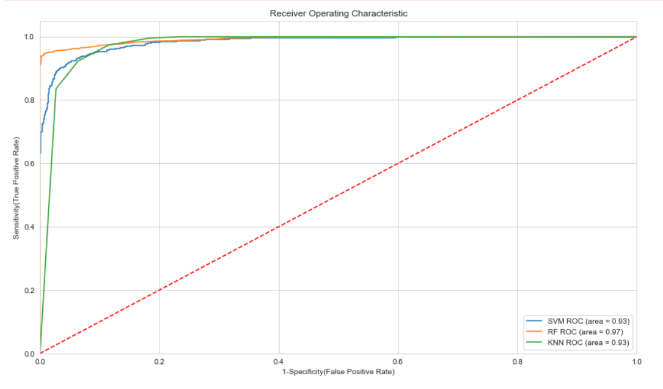
Accuracy = (TP+TN)/ (TP + TN + FN + FP), Precision = TP/ (TP+ FP), Recall = TP / (TP + FN),

F1 Score = 2.TP/ (2.TP+FP+FN), where TP is the number of true positives, FP is the number of false positives, FN is the number of false negatives and TN is the number of true negatives (Hernandez).

1. Results

In this work, the parameters accuracy, cross validation, precision, recall (sensitivity), and F1 score are computed to evaluate and compare the performance of the model- Random Forest (RF) and the two baselines- SVM and KNN for investigation of stroke detection. Table 1 displays the performance metrics for the three classifiers after optimisation.



All classifiers displayed satisfactory level of accuracy, especially the model- random forest and baseline- KNN which seemed to perform better than SVM overall. A plot of the performance measures is represented in Figure 1, as well as the Receiver Operating Characteristic (ROC) curve in Figure 2 where x-axis depicts Specificity (False Positive Rate) and y-axis is the Sensitivity (True Positive Rate) of the data.

1. Discussion

The confusion matrix was useful during experiments as is derives various measures including accuracy, recall (sensitivity), precision and F1 score along with the advanced measures such as ROC. Even though all models displayed good accuracy, other metrics were also considered as they provide with better overall performance measures. Precision shows the percentage of results that are relevant, whilst recall is the percentage of total relevant results correctly classified by your algorithm. The F1 Score is the harmonic average of Precision and Recall; it considers both false positives and false negatives, giving a better measure of incorrectly classified cases which tends to be better than accuracy. Random forest had the highest rates of every metric described and measured during the experiments, which means it is suitable to be applied for medical datasets as it will perform well to provide predictive results. KNN followed random forest with very similar results, indicating that predictions made by random forest were reliable. SVM had the lowest rates of metrics, yet it proved that random forest has good potential for good performance. Furthermore, the high ROC (Figure 2) and MCC show that random forest’s performance is the best at all classification thresholds and is the most statistically reliable, respectively.

**Conclusion 142 words**

The objective of this paper is to develop an approach based on machine learning which predicts any future possibility of patients suffering from stroke. The implementation of three different models, as random forest main along with SVM and KNN as baselines for comparison, to figure out the best model to predict if a patient is prone to suffer from stroke in the future. These classifiers were selected because they are well-known in building vulnerability predictors and have been used in a variety of similar research studies. From the evaluation, it is indicated that dataset used in this paper works well with random forest model as it outperforms the results obtained from the baselines. Random forest gave an accuracy of 97% along with higher rates of other evaluated performance metrics. Therefore, in the case of stroke prediction, random forest is a reliable model.

**References**

<https://www.kaggle.com/fedesoriano/stroke-prediction-dataset>