COMPS492F Machine Learning

Project Report (Topic 1)

**Stroke Prediction**

Student: CHIU Man Tsun (12317290)

**Introduction**

Stroke is the 2nd leading cause of death globally which is responsible for approximately 11% of total deaths (World Health Organization, 2020). According to the National Health Service (2020), stroke commonly happens when a blood clot blocks the flow of blood and oxygen to the brain. Although there is a correlation between stroke and some personal habits and physical conditions, not directly causality is found to bring stroke. Therefore, stroke prediction is an important way for studying stroke for preventing the appearance of stroke.

**Data Description**

The “Stroke Prediction Dataset” from Kaggle (2021) that used to predict stroke in this project provided personal habits and physical conditions of patients with or without stroke. This dataset contains 5110 observations with 12 attributes including the id, gender, age, hypertension history, heart disease history, marriage status, work type, residence type, average glucose level in blood, body mass index, smoking status, and stroke history of patients.

**Aim and Objectives**

This project is aimed to observe the correlation of different personal status and stroke events for preventing the occurrence. Finding the most suitable machine learning approach can effectively improve the studying of the cause of stroke. Therefore, this project will select a higher accuracy approach through comparison.

**Methodology**

Since this dataset is constituted to include categorical variables which contain nominal data, non-parametric and non-linear supervised learning algorithms K-Nearest Neighbors and Naïve Bayes Classifier are suitable for prediction in this project. Therefore, two approaches K-Nearest Neighbors and Naïve Bayes Classifier are selected to compare the predicting accuracy of stroke events. Moreover, two method K-fold cross validation and 7:3 train test split will be used to record the predicting accuracy of the above two machine learning approaches.

**Result**

By using K-fold cross validation (k=10) to compare the accuracy between K-Nearest Neighbors and Naïve Bayes Classifier, a higher cross validation score is obtained by K-Nearest Neighbors which was approximately 94% accuracy. Relatively, the cross validation scores obtained by Naïve Bayes Classifier was approximately 85% accuracy.

Furthermore, under a 7:3 train test split, K-Nearest Neighbors and Naïve Bayes Classifier respectively obtained an accuracy of approximately 95% (k=5) and approximately 87%. Although both machine learning approaches provided a high accuracy on predicting stroke, these two approaches show that K-Nearest Neighbors can give an approximately 10% higher accuracy than Naïve Bayes Classifier under any of the above methods on stroke prediction.

**Conclusion**

In conclusion, K-Nearest Neighbors provide a better predicting on stroke events with a high accuracy after comparison. It can be used to predict stroke events by recording some personal status for preventing any possible risk of stroke.

However, some limitations in this project may influence the accuracy of this project result. Only approximately 5% of patients in this dataset are having stroke history. Therefore, the prediction of stroke may still improve.

**References**

Fedesoriano. (2021, January 26). *Stroke Prediction Dataset*. Kaggle. https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.

National Health Service. (2020, November 16). Stroke. https://www.nhs.uk/conditions/stroke/

World Health Organization. (2020, December 9). *The top 10 causes of death*. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death