**Gintare Bucaite** [**gintare.bucaite@kcl.ac.uk**](mailto:gintare.bucaite@kcl.ac.uk)

**Challenge task for ElectronRx**

**Introduction**

Ischemic stroke occurs when arteries supplying blood to the brain become narrowed or blocked, usually by a clot, resulting in reduced blood flow. In the UK alone, there are more than 100,000 stroke incidences each year, making the stroke fourth biggest killer with economic burden of approximately £26 billion a year (1). The rate of first time strokes in people aged 45 and over is expected to increase by 59% in the next 20 years (2) and around two thirds of stroke survivors leave hospital with disability.

There is a number of modifiable risk factors associated with stroke, such as hypertension (high blood pressure), high blood cholesterol, diabetes (type 2), being overweight, smoking, alcohol consumption, drug use and lack of physical exercise (1, 3). In some studies, hypertension was regarded as the most important stroke factor (4).

**The dataset**

The cerebral vasoregulation in elderly with stroke (cves) [dataset](https://physionet.org/physiobank/database/cves/#background) contains multimodal data from a large study investigating the effects of ischemic stroke on cerebral vasoregulation.

This study compared 60 subjects who suffered strokes (6.1±4.9 years after stroke) to 60 control subjects. The control subjects were sex- and age- matched with no clinical history of stroke. The dataset largely contains data collected on blood pressure numerics.

Functional recovery after stroke depends upon integrity of cerebral vasoregulation. The data for this dataset was collected to test 3 hypotheses:

1. Older adults with ischemic stroke have impaired cerebral vasoregulation, rendering cerebral blood flow dependent on blood pressure.
2. Autonomic blood pressure control is impaired after stroke. Activities of daily living may induce hypotension, posing a risk of hypoperfusion.
3. The distribution of impaired vasoreactivity extends beyond the infarct region into surrounding gray and white matter, affecting other vascular territories.

**The task**

Here, the described dataset was employed for a machine learning task. General questions asked were:

1. Explore the data, generating visualisations for a subset of the feature columns. What is the clinical significance of these features? How are the features distributed?
2. Identify the clinical features which are most correlated with stroke?
3. Use these features to train a machine learning classifier, predicting stroke based on a subset of the available features. How does this model perform?

The task described here is a classification problem. Therefore, a supervised learning model was built to identify patients who belong to either (1) healthy control or (2) stroke group.

**Initial data analysis**

The dataset was cleaned up to include the entries where the patients have completed the visits. There was a large number of features of missing values, therefore the missing values were filling with the mean of the column. While this process may introduce bias, it allowed to maintain the richness of the dataset before the feature selection.

First, the dataset was explored to visualise the relationships between the high-risk factors associated with stroke and the target groups (stroke or control). For example, hypertension is reported as the most important risk factor (4) and it is reflected in Table 1 and Figure 1.

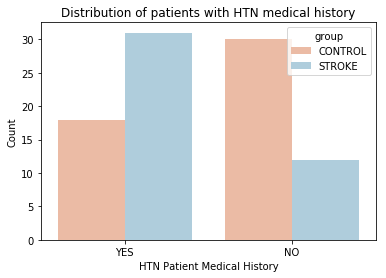


Table 1. Distribution of patients with hypertension (htn) medical history.

|  |  |  |  |
| --- | --- | --- | --- |
| group | htn\_patient\_medical\_history | count | %age |
| CONTROL | NO | 30 | 62.5% |
|  | YES | 18 | 37.5% |
| STROKE | NO | 12 | 28% |
|  | YES | 31 | 72% |

Figure 1. Distribution of patients with hypertension (htn) medical history between the control and stroke groups of patients.

Some high-risk factors, such as high cholesterol levels, were not so indicative of patients who suffered the stroke (Figure 2B), mostly because this group was heavily prescribed statins (Figure 2A), which can help lower low-density lipoprotein (LDL) cholesterol in the blood.



Figure 2. Distribution of control (red) and stroke (blue) patients receiving statins prescription (A) and the impact of statins prescription on the blood levels of cholesterol. Data shown as ±SD.

Due to a large number of features, visualisations of all of them using heatmap was not particularly informative. The number of features was reduced by identifying and removing features which were highly correlated together. This was performed for both numerical and categorical features.

**Model building**

LinearRegression() model was chosen to build a linear classifier model. Logistic regression can be used to predict binary outcomes (in this case: stroke or control). This algorithm was chosen over simple K-Nearest Neighbors (K-NN) because K-NN is non-parametric and all features have to be on the same scale and generally does not perform well with large number of variables.

Categorical variables (male/female, smoker/non-smoker) were transformed using pd.get\_dummies() function. This allowed incorporation of these features into the dataset.

For Linear regression model, the features should be independent of each other. Therefore, features which are highly correlated together were identified and removed from the dataset (correlation threshold was set to to 0.75).

Finally, a pipeline was built to test whether scaling the data prior to fitting the model improved the fitting. This proved to be successful, and final model had highest accuracy (0.9 or more).

**Future work**

The build model could be improved in several ways:

1. Reduce the number of features used in model building. In this model, features strongly correlated together were removed. Negatively correlated features should not be excluded. Additionally, features that do not add much to the model (do not correlate with target variable) should be removed for a more robust model.
2. Tune the hyperparameters for the model (logistic regression also has a regularization parameter: C). This could be performed by setting up a GridSearchCV to find the best C parameter. This would allow to avoid over or underfitting.
3. Perform an in-depth literature search and analysis to identify clinically relevant stroke risk factors and build models to test whether they are good indicators for patients with stroke.

**References**

1. Stroke Association. 2018. *State of the nation Stroke statistics - Februrary 2018*,.

2. Patel, A., V. Berdunov, D. King, Z. Quayyum, R. Wittenberg, and M. Knapp. 2017. *Current, future and avoidable costs of stroke in the UK, Executive summary Part 2*,.

3. Boehme, A. K., C. Esenwa, and M. S. V. Elkind. 2017. Stroke Risk Factors, Genetics, and Prevention. *Circ. Res.* 120: 472–495.

4. O’Donnell, M. J., D. Xavier, L. Liu, H. Zhang, S. L. Chin, P. Rao-Melacini, S. Rangarajan, S. Islam, P. Pais, M. J. McQueen, C. Mondo, A. Damasceno, P. Lopez-Jaramillo, G. J. Hankey, A. L. Dans, K. Yusoff, T. Truelsen, H.-C. Diener, R. L. Sacco, D. Ryglewicz, A. Czlonkowska, C. Weimar, X. Wang, and S. Yusuf. 2010. Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. *Lancet* 376: 112–123.