Case Study 3:
Predicting Central Neuropathic Pain (CNP)

GROUP- 59

Laharika Tutica	2797248T@student.gla.ac.uk
Yashika Rastogi	2802810R@student.gla.ac.uk
Sarika Kamble	2708819K@student.gla.ac.uk
Sanjana Jammigumpala	2761536J@student.gla.ac.uk

Introduction

In this case study, we are predicting Central Neuropathic Pain (CNP) in people with Spinal Cord Injury (SCI) from Electroencephalogram (EEG) data.

- CNP is pain in response to non-painful stimuli, episodic (electric shock), “pins and needles”, numbness
- There is currently no treatment, only prevention can be done using medicines which have strong side-effects.
- Here, we are predicting whether a patient is likely to develop pain is useful for selective treatment.

Dataset:

The data is pre-processed brain EEG data from SCI patients recorded while resting with eyes closed (EC) and eyes opened (EO). These tests were repeated 10 times on each participant.

- 48 electrodes recording electrical activity of the brain at 250 Hz
- 2 classes: subject will / will not develop neuropathic pain within 6 months
- 18 subjects/participants: 10 developed CNP and 8 didn’t develop CNP
- the data has already undergone some preprocessing
    - Signal denoising and normalization
    - Temporal segmentation
    - Frequency band power estimation
    - Normalization with respect to total band power
    - Features include normalized alpha, beta, theta band power while eyes closed, eyes opened, and taking the ratio of eo/ec.

Methods:

Feature Selection:

In case of dataset which has large number of features, there is a high possibility that some features are not relevant to generate the output. Such features increase the processing and evaluation time of the model and might cause the inaccuracy in the results. Thus, it is important to remove irrelevant features using feature selection methods. And then, model hyperparameter values are explicitly defined and tuned using the GridSearchCV in order to efficiently train the model.
Below are the feature selection methods used to simplify the dataset to predict Central Neuropathic Pain in the patients:

Filtering Methods: CHI Square

It is a statistical method which checks the dependence of output variable on the input variables. If there are variables of which output variable is independent those variables are considered to be irrelevant and thus dropped from the predicting model. Hence, final prediction is made according to the relevant features which helps in improving the prediction accuracy.

Wrapper Methods: Forward Feature Selection

It is an iterative method. Initially no feature is selected, but with each iteration a new feature is selected if it improves the performance of the model, till the iteration when addition of a new feature will not improve the performance.
While selecting the features for the prediction of Central Neuropathic Pain in the participants we have limited the iterations to 10 as forward feature selection is a time-consuming process.

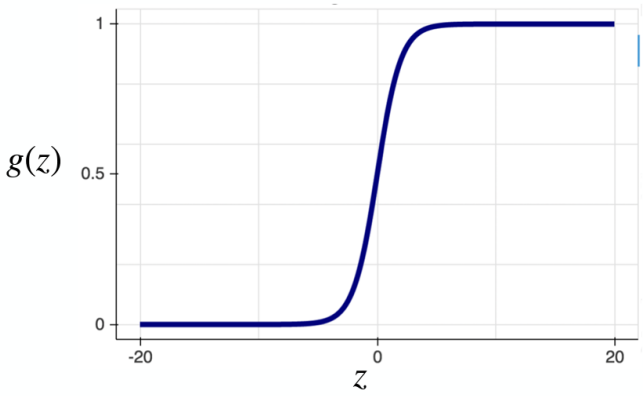
Differences between the filter and wrapper methods for feature selection are[1]:

- Filter methods are much faster compared to wrapper methods as they do not involve training the models.
- Filter methods measure the relevance of features by their correlation/ mutual information with dependent variable while wrapper methods measure the usefulness of a subset of features by actually training a model on it.
- Filter methods use statistical methods for evaluation of features while wrapper methods use cross validation.
- Feature selection using wrapper methods risk making the model more prone to overfitting.

CLASSIFIERS Used:

Logistic Regression:
It is one of the supervised learning algorithms used for Classification scenarios. Basically, it is used to predict if the event has occurred or not. Here, in the given dataset we are using it to identify if the patient will get the pain or not within the next 6 months.

Logistic Regression Model: hθ(x) = g(θ^Tx) where g(z) = 1/ (1 + e^−z).
Below is the graphical representation:

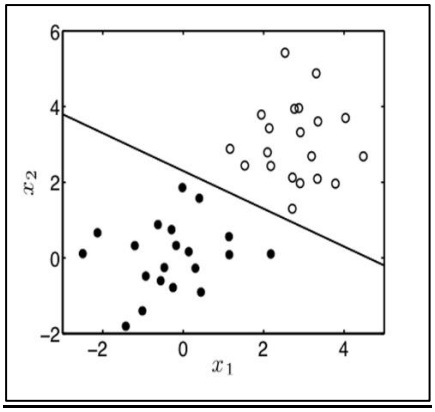


The important hyperparameters used to tune the model for the Logistic Regression classifier are:

- penalty: The values provided are ['l1','l2','elasticnet'] to compare the sparsity of the model.
- C: The values provided are [1,2,3,4,5,6,7,8,9,10,20,30,40,50]. The freedom of the model is said to increase with increasing values of C.
- max_iter: The values provided are [100,180,200,300] for the model to select the best number of iterations of the dataset for a better performance
- random_state: The values provided are [None,1,2,100] which assures that the training and testing split of data is chosen the same in every run thereby giving us the same result rather than choosing randomly.
- solver: The algorithms provided are ['lbfgs', 'newton-cg', 'liblinear', 'sag', 'saga'] and the best one is chosen to be used in the optimization problem.

Support Vector Machine:

It is also a supervised learning algorithm used for classification and regression scenarios. It generates a hyperplane which separates the datapoints plotted on n-dimensional space into classes.



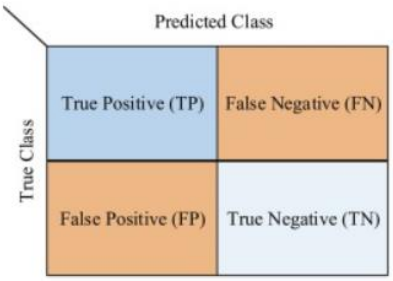
The hyperparameters used to tune the model for the SVM classifier are:

- C (Regularization parameter): The values provided are [0.1, 1, 10, 100] Basically, greater values of C reduce the chances of misclassification.
- gamma: The values provided are [1, 0.1, 0.01, 0.001, 0.0001,'scale','auto'] for the model to select accordingly. Generally, intermediate values prove to be a good decision for the model's accuracy.
- kernel: The values provided are ['rbf', 'poly', 'sigmoid'] and this crucial hyperparameter is used to improve the accuracy of the classification

Confusion Matrix :

It is a technique used to assess the performance of the classification algorithm.

- It is a table that is used in classification problems to assess where errors in the model were made.



- “True positive” for correctly predicted event values.
  “False positive” for incorrectly predicted event values.
  “True negative” for correctly predicted no-event values.
  “False negative” for incorrectly predicted no-event values.

Results

Experimental Setup:

The features have been selected according to the Chi- square and forward feature selection methods. After selecting the relevant features, each of the method has been validated using the classifiers Logistic Regression and Support Vector Machine.
The obtained accuracy score using the above-mentioned methods was then compared to the accuracy score of Baseline models (without feature selection).

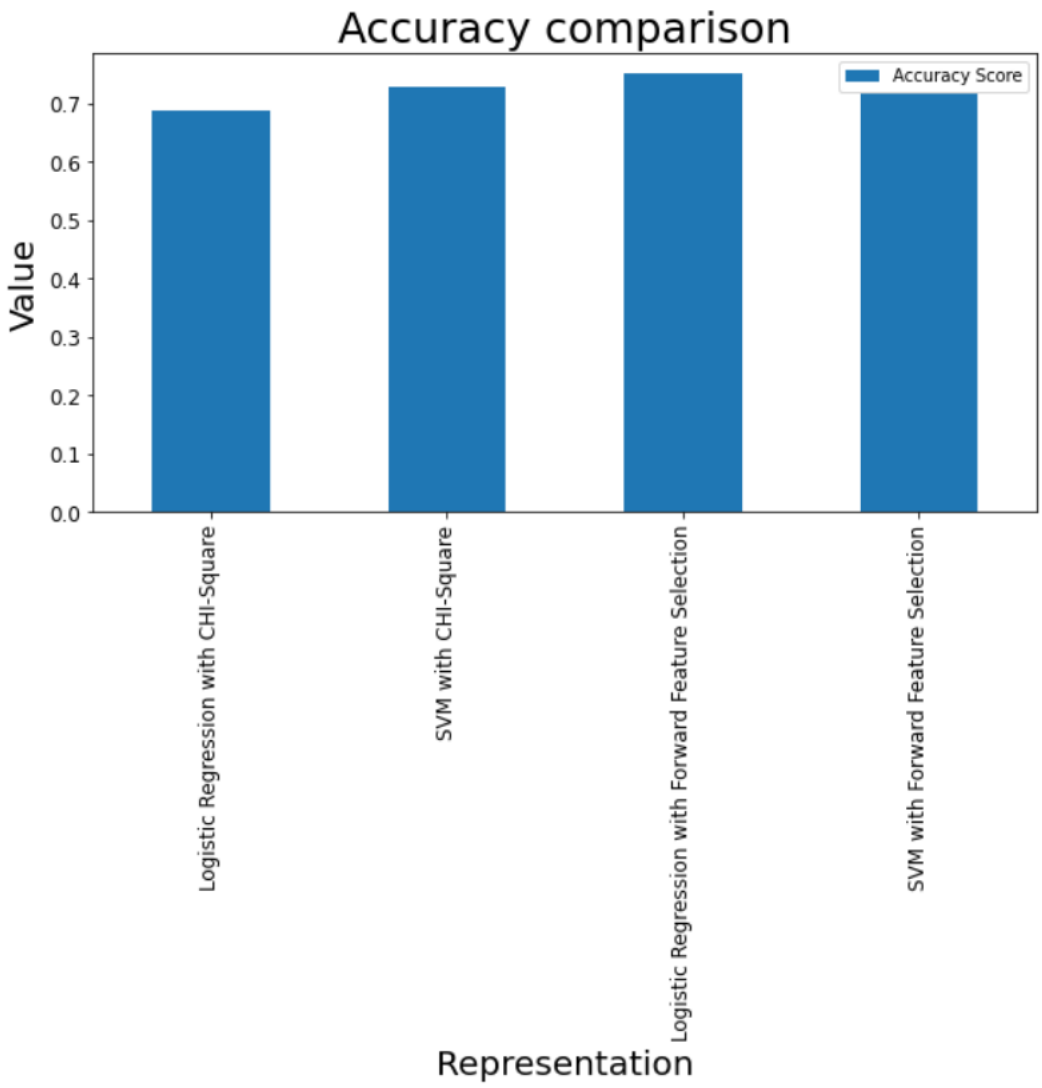
For training the models, the data has been split into 17:1 ratio with the observations of 17 participants (170 datapoints) as training data and the remaining one participant (10 datapoints) was used for testing the model according to the Leave-One-Group-Out cross validation technique and iterated it for 18 times.

Experiment Results:

Accuracy: Number of correct predictions over all predictions
Precision: Number of correct predictions
Recall: Number correctly predicted positive cases
F1-score: Harmonic mean of Precision and Recall

Below are the accuracy scores:

	CHI- Square		Forward Feature Selection		Baseline	
	Logistic Regression	SVM	Logistic Regression	SVM	Logistic Regression	SVM
Training Data	0.74	0.79	0.82	0.84	1.0	0.99
Testing Data	0.68	0.73	0.75	0.72	0.86	0.83



Classification report of Logistic Regression for CHI-Square feature selection:

	precision	recall	f1-score	support
0	0.75	0.86	0.80	7
1	0.50	1.00	0.67	4
2	0.88	0.64	0.74	11
3	0.50	0.80	0.62	5
4	0.62	0.38	0.48	13
5	0.62	0.62	0.62	8
6	0.75	0.55	0.63	11
7	0.88	0.88	0.88	8
8	0.50	0.44	0.47	9
9	0.75	0.75	0.75	8
micro avg	0.68	0.64	0.66	84
macro avg	0.68	0.69	0.66	84
weighted avg	0.69	0.64	0.65	84
samples avg	0.30	0.44	0.36	84

Classification report of SVM for CHI-Square feature selection:

	precision	recall	f1-score	support
0	0.75	0.75	0.75	8
1	0.62	0.83	0.71	6
2	0.88	0.64	0.74	11
3	0.62	1.00	0.77	5
4	0.75	0.43	0.55	14
5	0.88	0.64	0.74	11
6	0.88	0.54	0.67	13
7	1.00	0.89	0.94	9
8	0.88	0.50	0.64	14
9	0.88	0.88	0.88	8
micro avg	0.81	0.66	0.73	99
macro avg	0.81	0.71	0.74	99
weighted avg	0.83	0.66	0.72	99
samples avg	0.36	0.44	0.40	99

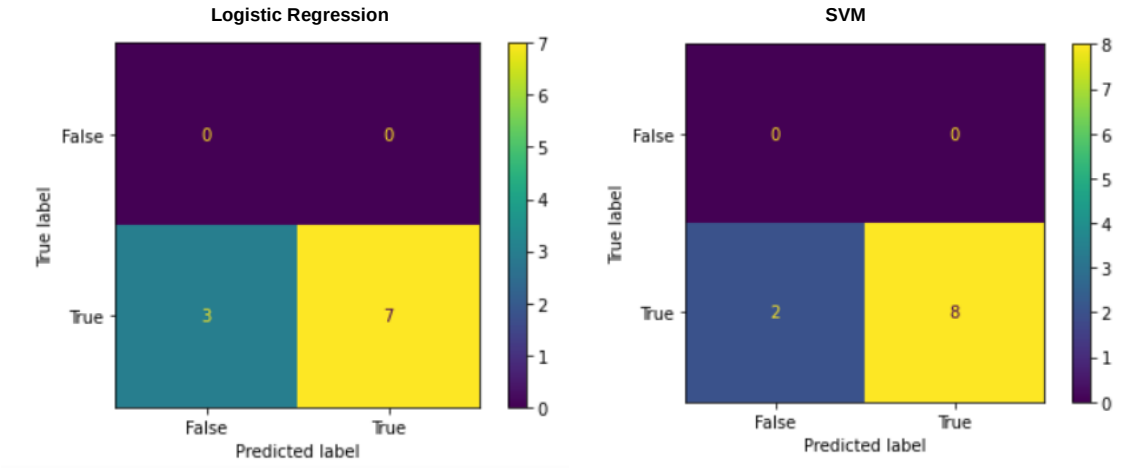
Classification report of Logistic Regression for forward feature selection:

	precision	recall	f1-score	support
0	0.38	1.00	0.55	3
1	0.62	1.00	0.77	5
2	0.38	0.50	0.43	6
3	0.75	1.00	0.86	6
4	0.38	0.60	0.46	5
5	0.75	1.00	0.86	6
6	1.00	0.57	0.73	14
7	0.75	1.00	0.86	6
8	0.50	0.50	0.50	8
9	0.88	0.88	0.88	8
micro avg	0.64	0.76	0.69	67
macro avg	0.64	0.80	0.69	67
weighted avg	0.70	0.76	0.70	67
samples avg	0.28	0.44	0.34	67

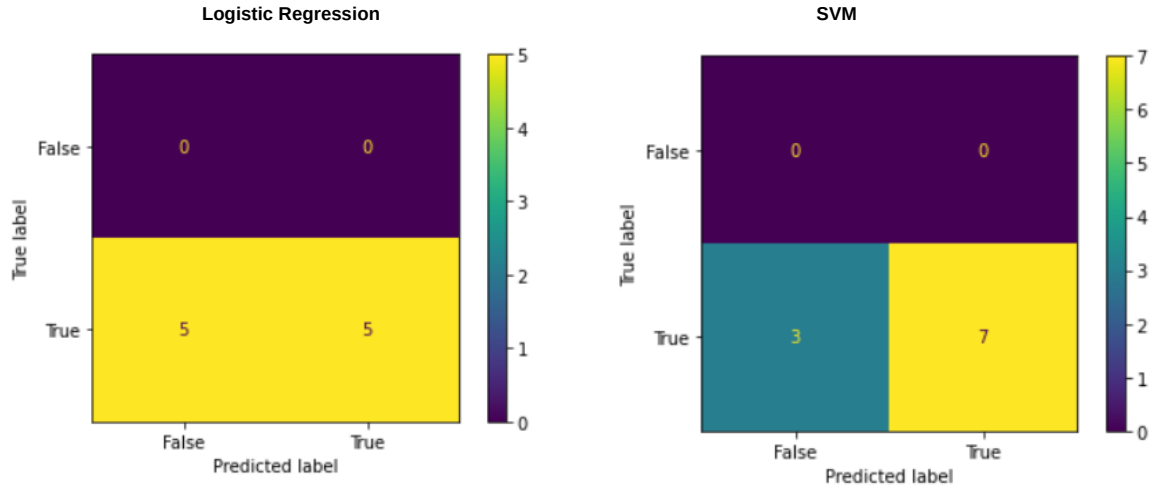
Classification report of SVM for forward feature selection:

	precision	recall	f1-score	support
0	0.38	1.00	0.55	3
1	0.62	1.00	0.77	5
2	0.38	0.50	0.43	6
3	0.75	1.00	0.86	6
4	0.38	0.60	0.46	5
5	0.75	1.00	0.86	6
6	1.00	0.57	0.73	14
7	0.75	1.00	0.86	6
8	0.50	0.50	0.50	8
9	0.88	0.88	0.88	8
micro avg	0.64	0.76	0.69	67
macro avg	0.64	0.80	0.69	67
weighted avg	0.70	0.76	0.70	67
samples avg	0.28	0.44	0.34	67

Confusion Matrix for Chi Square – Logistic Regression and SVM



Confusion Matrix for Forward Feature Selection – Logistic Regression and SVM



Conclusion:

- From the accuracy scores we can conclude that Baseline models have performed better due to the selection of all 432 features.
- According to the feature selection methods, forward feature selection has selected better features which lead to better accuracy scores when compared to CHI- square.
- In case of CHI- Square, SVM has been proved to be a better model whereas for the forward feature selection accuracy scores for the models are likely to be similar.

The applied models with feature selection methods couldn't perform very well due to the constraint of fewer datapoints and less related features with target variable (the correlation of most of the features with the target variable was found to be less than 50% when evaluated using Pearson correlation coefficient as a part of EDA-exploratory data analysis).

References

1. https://docs.google.com/presentation/d/1Rcia8EYBiiGI3mKe0CHUb8JJbYJTZab1hVbxoIkTg6I/edit#slide=id.g641413fc72_2_274
2. https://moodle.gla.ac.uk/pluginfile.php/5786651/mod_page/content/31/3_LogReg.pdf?time=1665486345253
3. https://moodle.gla.ac.uk/pluginfile.php/5786651/mod_page/content/31/5_SVM.pdf