**F21DL DATA MINING AND MACHINE LEARNING COURSEWORK 1 REPORT**

**Introduction:**

The study aims to detect emotions from a dataset of facial images, consisting of more than 35000 samples of 48x48 pixel grayscale images. In addition to reporting the interesting statistical characteristics of the data, it looks over the interrelation between various features and expressions which then lay as the groundwork for the Nearest Neighbor Classification.

**Steps 1 - 3: Data Conversion, Randomisation & Dealing with Computationonal Constraints**

The analysis has been carried out in R.

The CSV files were first loaded, split the pixel column to 2304 attributes, which was then partitioned to train and test sets and saved as separate files. The createDataPartition from caret package by default creates stratified random samples of the dataset. (Cran.r-project.org, 2018) The distribution of values in each column of the training dataset was visualised as shown in fig R1, R2 then preprocessed by eliminating the attributes whose standard deviation fell below 53.25 and scaled to unit variance.

**Step 4 : Data Dimensionality Reduction & Classification using knn**

For filtering attributes, PCA was performed using the prcomp from caret package on the scaled dataset (Fig R3) The PCA summary (fig R9) output says that the first 300 principal components account for nearly 95 % of the data. (Holland, 2018) Therefore six different models were trained using the data frames, created by choosing 1000 stratified random observations and principal components 1 to K as attributes where K {14,34, 68, 100, 140, 200, 300} as attributes. The root mean squared error was plotted as shown in fig R4. The least value was observed for the model that uses the first 68 PC which was then fitted for all observations in the training set. The test dataset was transformed and fed to the classifier. The maximum overall accuracy observed was more than 35% as seen in the confusion matrix in fig R5

The imbalance in the individual accuracy of classes is because the prevalence of class 3 is significantly higher and that of class 1 is dramatically lower than the rest, implying that the dataset is biased. For multi-class problems, the case where a single class is having a significantly more proportion than the average size of all other classes leads to overfitting. The converse, when multiple minority classes exist, random undersampling notably reduces the majority class performance. (Mosley, 2013) Therefore to evaluate this model, other performance metrics that don’t cloud our assessment with prevalence should be considered. (HarvardX Machine Learning, 2018) In the confusion matrix, the number of correct and incorrect predictions are summarised with count values and broken down by each class. (Brownlee, 2018)

**Steps 5 - 6: Feature Ranking usimg Absolute Correlation**

**Rosaria Silipo** mentions that “Data columns with very similar trends are also likely to carry very similar information. In this case, only one of them will suffice to feed the machine learning model. Here we calculate the correlation coefficient between numerical columns and between nominal columns” . (Seven Techniques for Data Dimensionality Reduction, 2018) According to Chaitanya Sagar, ”The idea is that those features which have a high correlation with the dependent variable are strong predictors when used in a model”. (FEATURE SELECTION TECHNIQUES WITH R, 2018)

In R, using Correlation function from SciViews package, the non-class features from each of the supplementary files were ranked in the order of absolute correlation to the emotion attribute. The top 10 features concerning each attribute are listed in **Table II.** Three other datasets were created, consisting of the top two, five and ten features from the list for each emotion, ultimately having fourteen, thirty-four and sixty-eight features respectively due to overlapping. The most accurate among the 3NN models trained on these datasets gave the confusion metrics as shown in the fig R6.

Looking at the confidence metrics, Sensitivity, which tells us what percentage ended up correctly classified for each emotion is higher for class 3(50 %). More the sensitivity, easier to find it. Implying that happy are painless to detect (Irizarry, 2018) whereas class 1 (9.1%) is harder to recognise. Specificity is greater for disgust emotion, which means that out of the 100% incorrect classes, 99.7 % was picked up as NOT class 1. This measure is least for class 3 (69.6 %). Balanced Accuracy, the one that measures the predictive quality for each class independently and aggregate (Mosley, 2013), says that nearly 65% of Surprises are predicted correctly. The confidence interval (CI) can be used to communicate how reliable the outcomes are. (En.wikipedia.org, 2018) Plot R7 compares CI of the three datasets. Broader the CI, more the reliability.

From the plot comparing the CV accuracy of the 3NN models using the top two five and ten features, ranked via absolute correlation and PCA, fig R5, we can infer that the overall accuracy improves with the number of essential attributes. Also, the feature ranking using PCA gives nearly 6 % higher results than High Correlation filter in this case.

**Conclusion**

* The emotion disgust is challenging to identify. The most wrongly predicted emotion is disgust.
* The top two features ranked by absolute correlation for each emotion is comparatively more reliable whereas the first five features from each are the least reliable among the three datasets.
* Feature ranking using absolute correlation in Task 5 and High Correlation Filter (Seven Techniques for Data Dimensionality Reduction, 2018) Feature selection performed for Task five and six are part of Dimensionality Reduction, which is a powerful technique beneficial for exploratory data analysis. (Irizarry, 2018)
* Failure to randomise the datasets used for NN classifier can cause the outcomes to seem either larger or smaller than they are. (Kunz and Oxman, 1998)
* The cross-correlation of a signal/image with itself is termed as Autocorrelation. Intuitively, it is the similarity between observations as a function of the time lag between them. (En.wikipedia.org, 2018) If there is a cross-correlation between the features within an image, then the autocorrelation coefficient can be used as discriminant features (Popovici and Thiran, 2004), implying that these can characterise or separate two or more classes. (En.wikipedia.org, 2018)

## **References**

Brownlee, J. (2018). What is a Confusion Matrix in Machine Learning? [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/confusion-matrix-machine-learning [Accessed 18 Oct. 2018].

Cran.r-project.org. (2018). A Short Introduction to the caret Package. [online] Available at: https://cran.r-project.org/web/packages/caret/vignettes/caret.html [Accessed 18 Oct. 2018].

En.wikipedia.org. (2018). Confidence interval. [online] Available at: https://en.wikipedia.org/wiki/Confidence\_interval [Accessed 18 Oct. 2018].

En.wikipedia.org. (2018). Cross-correlation. [online] Available at: https://en.wikipedia.org/wiki/Cross-correlation [Accessed 18 Oct. 2018].

FEATURE SELECTION TECHNIQUES WITH R. (2018). [Blog] Dataaspirant. Available at: http://dataaspirant.com/2018/01/15/feature-selection-techniques-r/ [Accessed 18 Oct. 2018].

HarvardX Machine Learning. (2018). .

Holland, S. (2018). Data Analysis in the Geosciences. [online] Strata.uga.edu. Available at: http://strata.uga.edu/8370/lecturenotes/principalComponents.html [Accessed 18 Oct. 2018].

Irizarry, R. (2018). Introduction to Data Science. [online] Rafalab.github.io. Available at: https://rafalab.github.io/dsbook/ [Accessed 18 Oct. 2018].

Irizarry, R. (2018). Introduction to Data Science. [online] Rafalab.github.io. Available at: https://rafalab.github.io/dsbook/dimension-reduction.html [Accessed 18 Oct. 2018].

Kunz, R. and Oxman, A. (1998). The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ, 317(7167), pp.1185-1190.

Mosley, L. (2013). A balanced approach to the multi-class imbalance problem. Iowa State University Digital Repository.

Popovici, V. and Thiran, J. (2004). Pattern recognition using higher-order local autocorrelation coefficients. Pattern Recognition Letters, 25(10), pp.1107-1113.

Seven Techniques for Data Dimensionality Reduction. (2018). [Blog] KDNuggets. Available at: https://www.kdnuggets.com/2015/05/7-methods-data-dimensionality-reduction.html [Accessed 18 Oct. 2018].

En.wikipedia.org. (2018). Linear discriminant analysis. [online] Available at: https://en.wikipedia.org/wiki/Linear\_discriminant\_analysis [Accessed 18 Oct. 2018].