Declaration on Plagiarism

This form must be filled in and completed by the student(s) submitting an assignment

Name:	Sunil Jagap & Pavan Kirageri
Student Number:	20211080 & 21262357
Program:	Master in Computing (Data Analytics)
Module Code:	CA682
Assignment Title:	Data Visualisation
Submission Date:	26 Nov 2021
Module Coordinator:	Dr. Suzanne Little

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offenses in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I have read and understood the Assignment Regulations. I have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

I have read and understood the referencing guidelines found at h ttp://www.dcu.ie/info/regulations/plagiarism.shtml, h ttps://www4.dcu.ie/students/az/plagiarism and/or recommended in the assignment guidelines

Name: Sunil Jagap and Pavan Kirageri Date:26/11/2021

Project Title: Analysing Genetic makeup of different types of ALS patients.

Abstract

Amyotrophic Lateral Sclerosis (ALS) is a fatal neurological illness that affects one out of every 500,000 persons. The most famous case of ALS is Stephen hawking. However, awareness of the disease was limited and many of the cases go undiagnosed. A massive awareness campaign known as the Ice bucket challenge was undertaken in mid-2014 which raised funds and created awareness among the general population about this disease. There are no known medications to cure ALS and very limited therapies and drugs are available to help patients. Patients lose voluntary muscular control over time, resulting in paralysis, difficulty speaking, eating, and, eventually, breathing difficulties. Moreover, 60% of ALS patients die after three years of diagnosis, and 90% die within ten years. There is, however, cause for optimism.

ALS is an autoimmune disease where the body starts degenerating motor neurons. Hence analyzing genetic data is the most important task. Modern technologies like Biotechnologies are quickly changing the way we learn about the brain in health and illness by producing new sources of data. Patients with ALS and healthy controls have produced an unparalleled amount of clinical and biochemical data. We need to analyze the same biochemical data to gain some insights into the root cause of this enigmatic disease.

Dataset

Data was posted as a part of the end ALS challenge in Kaggle. The dataset consisted of a collection of Genomics, transcriptomics, and clinical data of 134 people. Three different transcriptomics data were available for examination.

1. Patients with bulbar onset (disease starts with difficulty in speaking and eating) vs limb (starts with difficulty in controlling hand and leg moments), 116 patients and 53000 genes (approx) in which 31 with onset on bulbar regions and 85 are with onset in limb region.
2. Ctrl vs Case (genetic makeup of people with ALS and healthy people), 169 patients and 53000 genes approx in which 32 are controls and 137 ALS cases
3. Median Scores: ALSFRS scores [1] which is a way to score disease progression, 92 patients, 45 with Scores less than median and 46 with higher scores.

The data is anonymized, i.e the names of the patients were not published. The data is open for public use and was used as a part of the Kaggle challenge.

We used another dataset Ensemble bio mart dataset to annotate the genes, to remove pseudogenes that are not important for our analysis.

Data Exploration, Processing, Cleaning, and/or Integration

We examine the transcriptomics portion of the data and use PCA[2] to reduce the dimensionality and visualize it (Principal Component Analysis). (Data from transcriptomics - each gene expression for each patient is shown.) The biggest dataset has 53861 genes and 163 patients.

That is a common approach when dealing with transcriptomics data; occasionally, we get lucky and observe certain clusters that match the goal variable or have biological significance. That is not the case with that data, unfortunately, PCA analysis didn’t give us as many insights as per expectations.

In the next method, we tried feature selection methods based on fisher's scores or laplacian scores. In the final phase, we'll analyze each fold's forecasts.

Over 53,000 genes are found in each RNA-seq sample in the DEseq2 folder. In order to accurately forecast a patient's illness condition, we must do some type of gene filtering and feature selection in our model. Some genes may be ruled out as uninformative before employing any of these approaches. "Pseudogenes," which are nonfunctional DNA segments that look like functional genes, are one example of these genes. It's possible that certain pseudogenes in the datasets appear to be connected with our target variable through coincidence, but they have no biological significance in predicting ALS. As a result, we may want to delete these genes from our model before using any feature selection approaches. To obtain a list of gene ids that are recognized as pseudogenes.

Visualization

The data consisted of transcriptome information of various patients. We analyzed the bulbar vs limb and ctrl vs case dataset. We ran a PCA analysis on these 2 datasets to find similarities and differences between the transcriptome data of these patients. In PCA distance between two points represents the magnitude and angle represents correlation between them. Scatterplots display enormous amounts of data and make it simple to identify relationships and clustering effects. So We plotted a scatter plot for this purpose so that we can identify easily whether there is any visible distinction between two different groups. Unfortunately, this didn't give us any productive insights into the distinction between groups so we used different analysis techniques to get variable genes between two different groups.

Since PCA is an unsupervised learning method and fisher's LDA and Laplace feature selection method is supervised learning. Both of these methods are used for feature selection. We used the fisher scoring algorithm[4] and laplacian[5] algorithm for feature selection. We tried running both feature selection methods on known genes that might cause ALS as per medical researchers on our dataset and the same method on all genes available in the data.

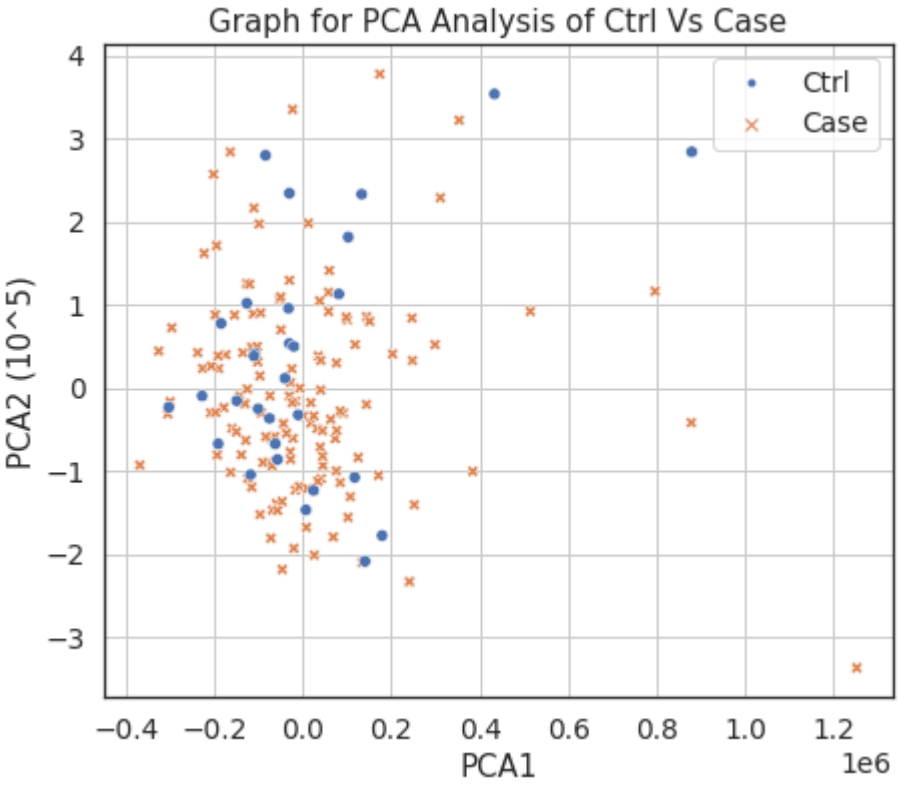

Figure-1

Figure-1 shows PCA-1 and PCA-2 scores for the control vs case dataset where

1. Blue round -> control patients (patients without ALS)
2. Orange stars -> ALS patients
3. The X-axis represents PCA-1 and The y-axis represents PCA-2 (10^5)

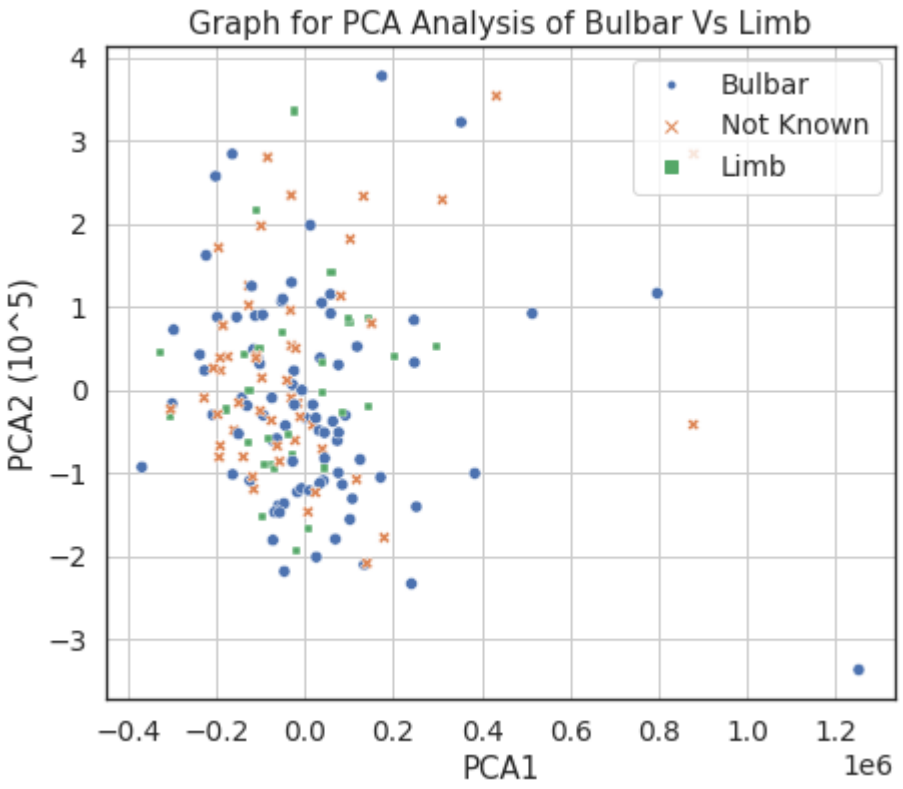

Figure-2

Figure-2 shows PCA-1 and PCA-2 for bulbar vs limb patients

1. Blue-round -> Patients with bulbar onset
2. orange x -> Point of origin of disease not known
3. Green Square -> Patients with limb onset
4. The X-axis represents PCA-1 and the y-axis represents PCA-2 (10^5)

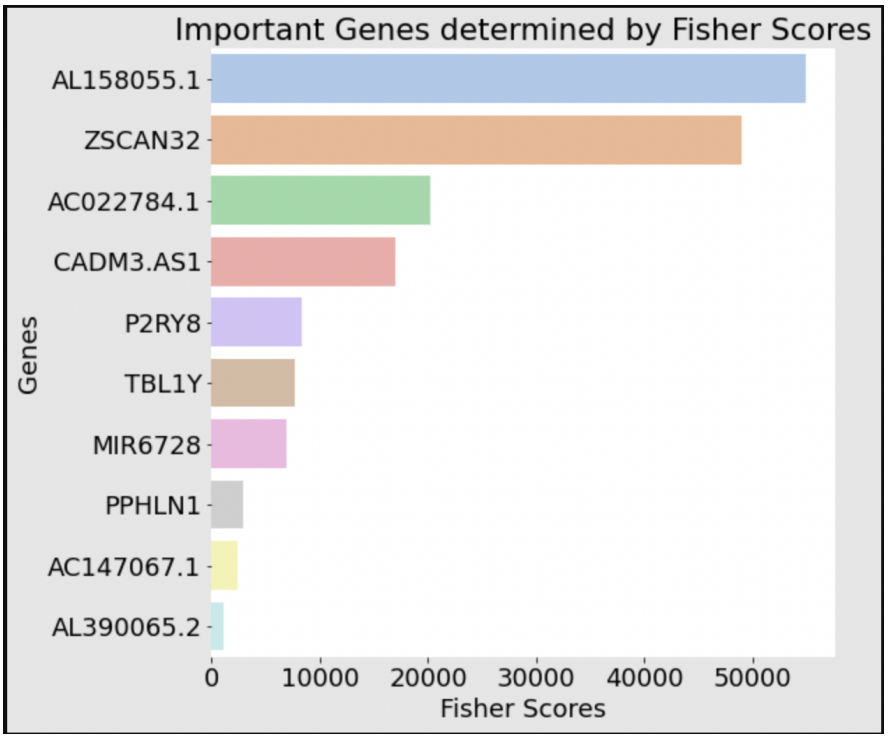

Figure-3

Figure-3 Is a bar chart highlighting important genes as determined by fisher scores for control vs case dataset

1. X-axis represents fisher scores
2. Y-axis -> Type of genes

Only genes that have fisher score of more than 1000 are represented in this graph

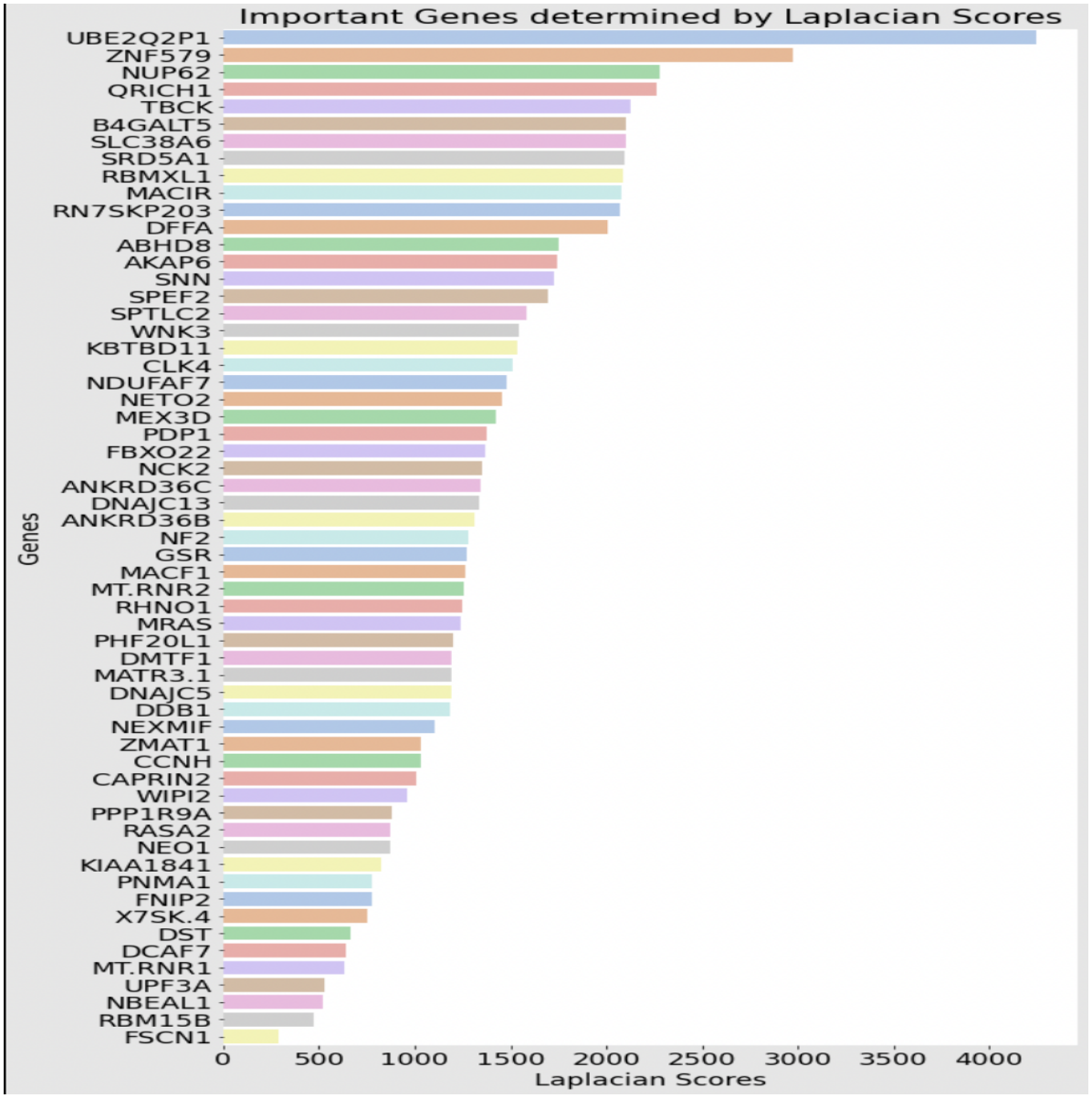

Figure-4

Figure-4 Represents laplacian scores for important genes as determined by laplacian scores for control vs case dataset

1. X -axis -> laplacian scores
2. Y axis -> Type of genes

Only genes that have laplacian scores of more than 250 are represented.

Tools used:-

- Jupyter Notebook (Python) for Exploratory Data Analysis
- Python Library Seaborn and matplotlib
- Kaggle for dataset and data cleaning

Conclusion

In this project, we have studied data preprocessing techniques like PCA, laplacian, and fisher

analysis along with different data visualization libraries and techniques like matplotlib, seaborn. Pandas were used for easy data manipulation. Along with an understanding of genetics and transcriptomics to preprocess and understand the dataset.
Since PCA is unsupervised learning it is hard to form clusters from genetic information provided. Unfortunately, PCA didn't really help us in gaining any insights into the dataset. However, fishers and laplacian techniques extracted some of the RNA transcripts that might be potential causes of the disease and helped us understand the difference between healthy and ALS patients. But fisher and laplacian scores for transcriptomics data were not consistent. So more research has to be done to validate the results. Complex methods like artificial neural networks, reinforcement learning, or genetic algorithms can help solve the problem.

References

[1] *Sralab.org*. Available at: https://www.sralab.org/sites/default/files/2017-07/PMandR_ALSRatingScale033111.pdf (Accessed: November 25, 2021).

[2] Richardson, M. (no date) *Principal Component Analysis*, *Cuni.cz*. Available at: http://aurora.troja.mff.cuni.cz/nemec/idl/09bonus/pca.pdf.

[3] He, H., Tian, C., Jin, G. *et al.* Principal component analysis and Fisher discriminant analysis of environmental and ecological quality, and the impacts of coal mining in an environmentally sensitive area. *Environ Monit Assess* 192, 207 (2020). https://doi.org/10.1007/s10661-020-8170-0

[4]Li, C. and Wang, B. (no date) *Fisher Linear Discriminant Analysis*, *Northeastern.edu*. Available at: https://www.ccis.northeastern.edu/home/vip/teach/MLcourse/5_features_dimensions/lecture_notes/LDA/LDA.pdf

[5]He, X., Cai, D. and Niyogi, P. (no date) *Laplacian Score for Feature Selection*, *Edu.cn*. Available at: http://www.cad.zju.edu.cn/home/dengcai/Publication/Conference/2005_NIPS_LaplacianScore.pdf.

[6]Kapri, A. (2020, February 17). *PCA vs LDA vs T-SNE — Let's Understand the difference between them!* Analytics Vidhya. https://medium.com/analytics-vidhya/pca-vs-lda-vs-t-sne-lets-understand-the-difference-between-them-22fa6b9be9d0

[7]rwilliams (2021) *ALS Challenge Task 1*, *Kaggle.com*. Kaggle. Available at: https://www.kaggle.com/rwilliams7653/als-challenge-task-1

[8] alexandervc (2021) *ALS Transcriptomics Visualizations via DimReduct*, *Kaggle.com*. Kaggle. Available at: https://www.kaggle.com/alexandervc/als-transcriptomics-visualizations-via-dimreduct