Sweta Rauniyar ‘21
Professor Douglas Luman
CMPSC 480
Code Project

A Profiling Method in the Early Detection of Illness

Introduction:

This project originates from the field of Bioinformatics and concerns the study of genes and their levels of activity in producing proteins that may likely be associated with disease-onset and cancer-initiation. We note that the study of gene expression data will provide important insight into which genes are responsible for the onset of particular types of disease. The knowledge of which particular genes are present and are being expressed during the initiation of disease will enable us to build a type of primitive profiling system which may be used in conjunction with other diagnostics for early detection. In order to make it convenient to work with numerous sets of data, we implemented the techniques of Machine Learning to help us detect signals within the data and eventually make decisions with minimal human intervention. Among the wide variety of algorithms and statistical models available in Machine Learning, that contribute to the art of pattern recognition, we used the Hierarchical Clustering Algorithm that works on the basic concept of a hierarchical tree that represents data in the form of nodes and each node in the tree contains similar data.

Method

For this project, we accessed data from the National Institute’s (NCI’s) Genomic Data Commons which is a data sharing platform that allows users to share and download genomic and cancer data for analysis. With the convenience for users to add data, the GDC acts as a powerful tool as researchers have access to a diverse and large volume of clinical data that can be analysed and possibly lead to findings that could benefit the patients.

We started our research by working with two data sets containing breast cancer and blood cancer data respectively. Our gene expression data consists of a list of genes accompanied by the protein production rates from individual cancer patients. The rows within the data consists of the names of the genes while the columns include the numerical values corresponding to the production of proteins. We chose to work with a small volume of data that includes information from less than three hundred genes. The reason we chose to do so was because implementing the computer program was our next step in the project and we wanted to make sure that the program runs correctly. The efficiency and correctness of a new program would be easier to test with a small set of data rather than a bigger one.

Implementation

Our program is written in Python which is a high-level programming language. Python provides a number of packages that help with data handling. However, the basic essence of our program lies within the realms of Machine Learning which is a subset of Artificial Intelligence.

Machine Learning involves making a computer perform a certain set of tasks without having the humans to do explicit programming. In other words, it is based on the idea that systems can learn from the data using automation, identify signals within the data, and make decisions with minimal human intervention. There is a magnificent amount of data in the world and it is beyond human capacity to handle all of this data and use it for better results.

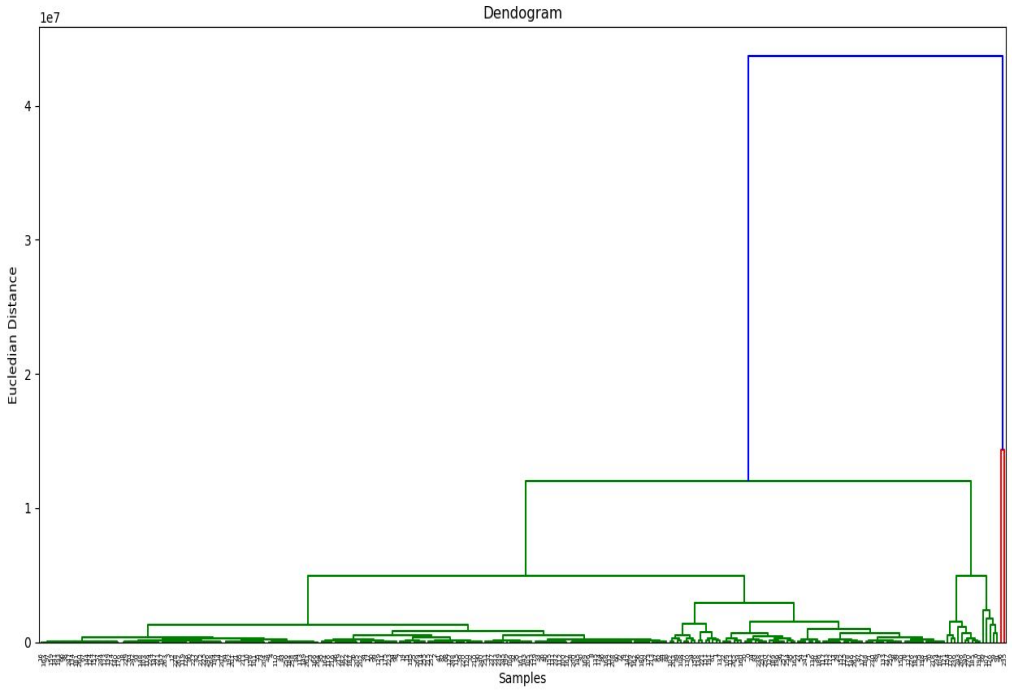
Machine learning processes has helped developers to create automated systems for various industries that has helped them to provide fast and efficient services for their customers. A few examples of such applications include online recommendation systems at Amazon or Netflix, tracking location and costs for modern transportation services like Uber, automated email responses etcetera. Although these are applications from different areas, what remains common between them is that these systems pay attention to the previous activity of the users and use that information to make relevant future decisions. While all of the above examples provide comfort to the users, we aim to use the concept of Machine Learning in the field of medicine by studying the genes and the productivity of the proteins that could possibly lead us to valuable findings. Eventually, these findings can be helpful for researchers to find measures that could benefit patients.

Machine learning uses algorithms to learn from the data set and make useful predictions. The *Scikit-Learn* package in Python provides access to different algorithms that includes supervised and unsupervised learning. Supervised learning is a machine-learning algorithm where we infer from a labelled training data. On the other hand, unsupervised learning deals with data with no labeled outputs. It helps in finding the unknown patterns within the data and that is why we will be using this approach because our data is not labelled and we aim to find the hidden patterns within the data.

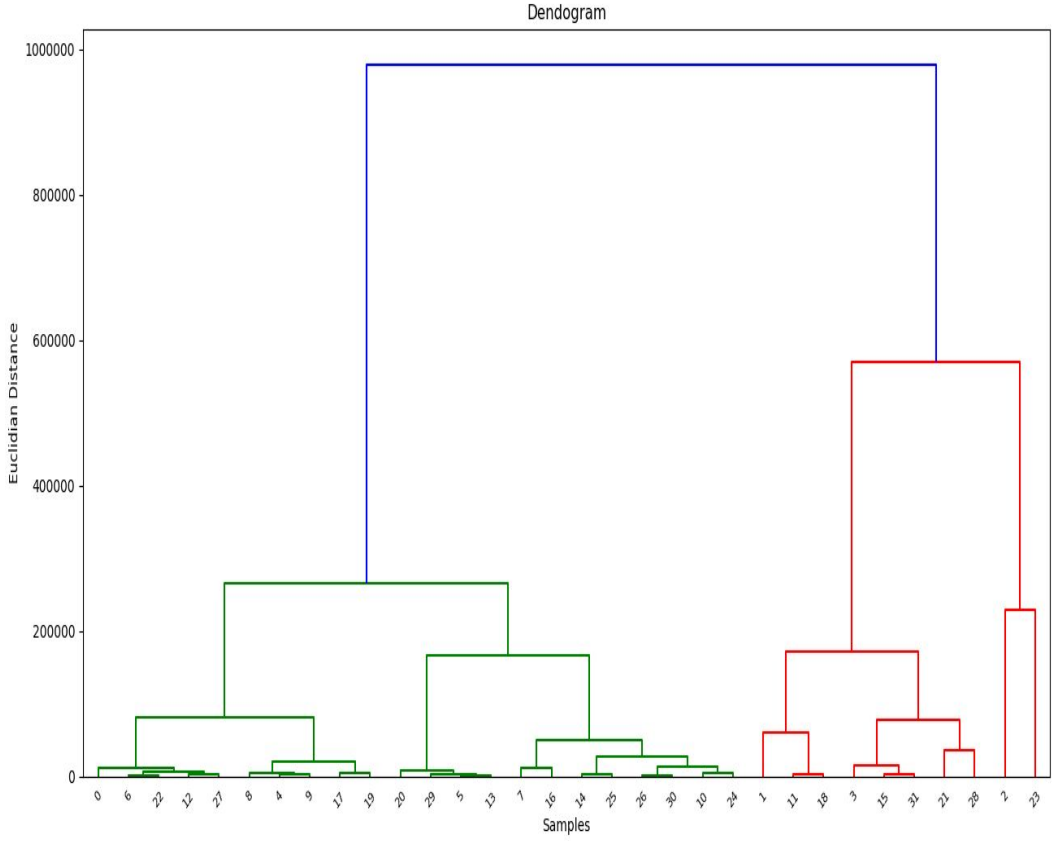
One of the popular methods in unsupervised learning is Cluster Analysis. This involves dividing the data that are similar to each other in individual groups. The formed clusters will help us find structures within the data however the question: “Why these structures exist?”, is up to the medical researchers to find out. In Cluster Analysis, there are various methods to perform the clustering process. For our project, we leaned towards the Agglomerative Hierarchical Clustering algorithm. This kind of clustering involves creating a tree-like diagram called dendrogram where all the data is classified in the form of groups or nodes.

Results

The dendrogram helped us to create an obvious visual representation which portrayed the merges between the genes. The genes that were merged together had similar protein activity and the interconnected nodes showed how the merge between the different genes are possible for a out-break of a disease. The following figure is a dendrogram that was created with the data from Breast Cancer. The x-axis includes all the gene samples from different patients while the y-axis includes the euclidean distance calculated with the values of the protein activity from all the genes.



At the bottom of the figure, we can see how the different points have been formed into a number of clusters. Each of these clusters are merged as they have a similar centroid distance which is calculated by the Euclidean distance. It was somewhat challenging for us to find details within the formed clusters as it was not obvious from the above graph itself. In order to resolve this, we plotted another graph by reducing the number of data points so that we could have a closer look at the individual clusters. The plotted data with the subclusters is shown below.



On creating this dendrogram for breast cancer, we can assume what our hierarchical tree would look like if we were to plot the different gene samples that we had used for Breast Cancer.

Conclusion

With further research with data on other different kinds of cancer concerning the same genes, we could see if our hierarchical trees have any similarity between each other. In case of the presence of any kind of similarity, we can detect the signals within the data that there are certain genes that can contribute to the possible out-break of a disease. However, medical data is very sensitive and every patient has a unique medical profile and we will have to work with diverse data sets in order to conclude with certain patterns. We are using Machine Learning to determine the inherent signals in the data. Medical profiles are created from signals which could help medicine to recognize similar occurrences in other samples of data. The amount of medical data we have in the world is truly mind-boggling. Our goal was to use this available data as an important tool that can help us gain an important insight and eventually use this information in practicing relevant medical research