1. **INTRODUCTION**

Over the years risk factors causing Leukemia have been studied extensively for the ailment that affects hundreds of thousands of people globally. These risk factors have been whittled down to genetic factors including genetic syndromes such as down syndrome, Li Fraumeni syndrome etc. Other factors include environmental factors such as radiation, and lifestyle related risk factors such as smoking. In this assignment, we’ll be focusing on the genetic components of this disease as I describe the resource use data obtained on gene expressions of different leukemia cases.

What I hope to achieve in this data analysis is to figure out which gene expressions have the most impact and are the highest risk factors for different types of leukemia. Due to the small number of samples the report will focus mainly on gain insights on the data rather than prediction.

1. **DATA SUMMARY**

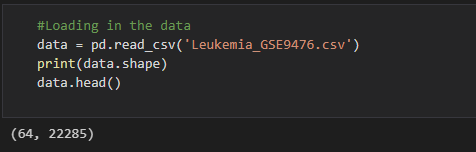
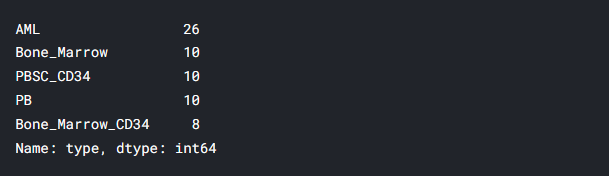
To begin my analysis, I’ll check the basic structure of the data presented. The data has 64 rows and 22284 columns. The data includes 5 classes of leukemia, 26 cases of Acute Myelogenous Leukemia cases (AML) 10 cases of PBSC\_CD34 gene expression cases, 10 cases of Bone Marrow Leukemia cases, 10 cases of PB expression cases, and 8 cases of Bone Marrow CD34 expression cases.

Figure : Value counts of the data

Figure :Shape of the data

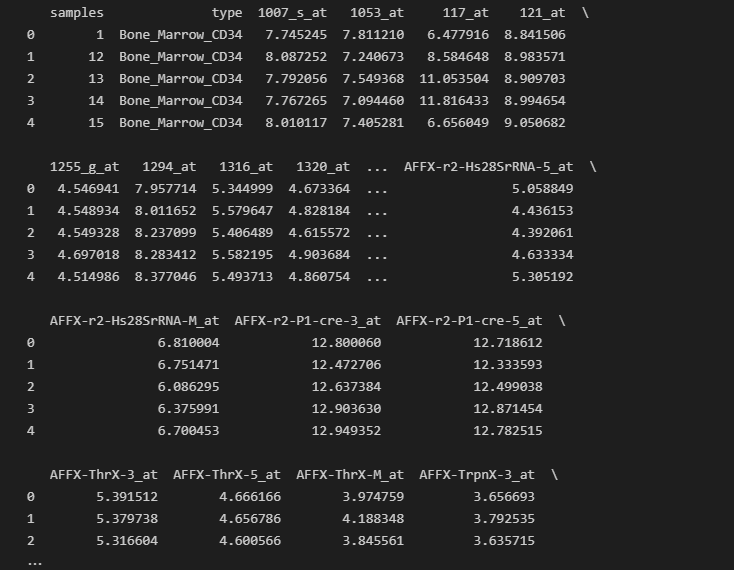


Figure : Sample data

Given the nature of the data classification models would be the most appropriate models to apply to the data to gain some required insights. One challenge that is going to be faced is overfitting seeing as the data has way more features than samples. But regardless, amazing insights can be made by what’s available.

1. **DATA EXPLORATION AND FEATURE ENGINEERING**

To start off we need to convert the labels into something that can be fed into a classification algorithm. The classification algorithm cannot read or interpret text. To solve this, we’ll be using a label encoder to convert the labels to integer categories that are more appropriate for a mathematical model.

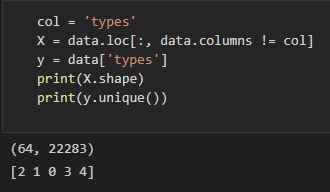


Figure : Classes after label encoding

Once the labels have been encoded, we need to scale the data in order to make sure the ranges are between 0 and 1. This will also assist the models in processing the data faster. The data itself had most values between one and zero so the alterations were not drastic.

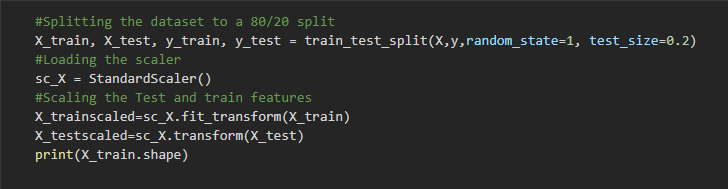


Figure : Scaling the data

The samples column is also removed as it has the least effect on the target and it doesn’t contribute to identify types of leukemia.

1. **SUMMARY OF MODELS**

The first model was the multilayer perceptron implemented using Scikit Learn. The MLP Classifier is a group of feedforward artificial neural networks that are composed of multiple layers or perceptrons. They include at least three layers, an input layer, hidden layers and an output layer. In our case we will fit our data into a model into a standard Scikit Learn MLP Classifier. The classifier will have 4 hidden layers each layer halving the number of nodes of the previous layer starting with 256 nodes. The activation function of the classifier will be ReLu. After fitting the data and training the model the classifier has a perfect accuracy score which could either be a sign of overfitting as we anticipated with the small number of samples, or the efficiency of the model

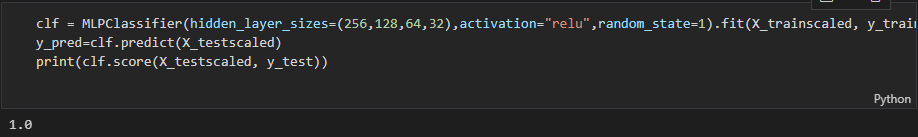


Figure : MLP Classifier

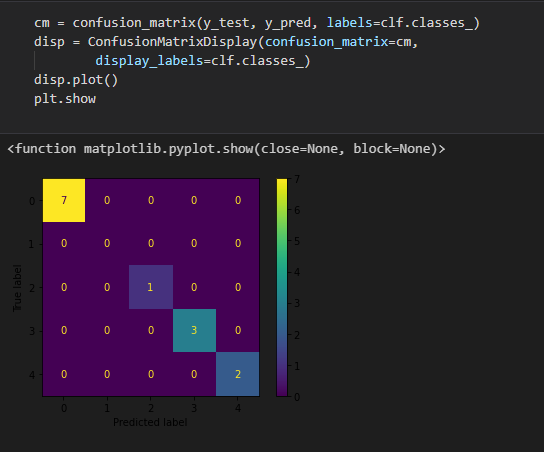


Figure : Confusion Matrix on the MLP Classifier

The second model was a Convolution Neural Network implemented using TensorFlow. Using a Conv1D model, we fit the data on the batch of models. We use a ReLu activation function by invoking the Leaky ReLu method. We will also invoke the MaxPooling activation functions as well as Flattening the data to gain more insights from the data.