**Gene Expression Classification**

**CSCI417**

* **Habiba Mohamed Ibrahim 19105386**
* **Malak Mohamed**
* **Manar Mohamed**
* **Renada Amr**
* **Wafa mohsen**

2022

**Table of contents:**

* Problem statement.
* Related work
* Model architecture
* Evaluation results and strategies.
* GitHub link.
* **Problem statement.**

Leukemia affects over 44000 people in the United States alone each year, and it has one of the highest mortality rates among all types of cancer. Leukemia is a type of blood cancer. It is one of a group of diseases known collectively as hematological malignancies. If the disease affects the lymphocyte-forming marrow cells, it is referred to as lymphocytic or lymphoblastic. T-cell acute lymphoblastic leukemia (T-cell ALL) affects 15% of lymphoblastic leukemia patients, while acute B-cell lymphoblastic leukemia (B-cell ALL) affects 85% of ALL patients.

If the acute disease affects the bone marrow cells that produce red blood cells, white blood cells, or platelets, it is referred to as acute myelogenous or acute myeloid leukemia (AML). There is an urgent need to detect and classify the disease at its inception. Furthermore, accurate subtype classification is critical because treatment protocols differ significantly between subtypes.

Leukemia cells have traditionally been classified based on their morphological appearance. To distinguish between tumor cells, highly skilled are required. The procedure can be time-consuming, labor-intensive, and costly. It is worth noting that the technique is not completely effective even when all of the resources required are available. Cells can have morphologically similar appearances.

The limitations of the traditional technique necessitated the identification of additional parameters that could be used as a basis for cell categorization. Data on gene expression may be useful for sub-classification studies. DNA microarrays have been critical in monitoring gene expression data from thousands of genes at once. It is possible to compare the levels of gene expression in normal cells and abnormal cells. The study of expression levels can provide useful insights for making classification decisions based on the gene signature of the cells under consideration.

Despite advances in research and classification of cancer over the past 30 years, there is still no reliable method for class discovery or classifying tumors into recognized classes (class prediction). In order to demonstrate an accurate method of cancer classification based on gene expression monitoring by DNA microarrays, our project describes and uses a test case of human acute leukemia. A class discovery approach classifies acute lymphoblastic leukemia (ALL) and acute myeloid leukemia(AML) based on gene expression monitoring by DNA microarray and without any prior knowledge of these classes.

We have used more than 10 models to classify leukemia patients into ALL or MLL and compare their accuracy

* **Related work**

Microarray technologies enable the measurement of the molecular signatures of cancer

cells. These data allow different types of analyses, such as (1) the identiﬁcation of

differentially expressed genes [5], which could indicate possible gene targets for more

detailed molecular studies or drug treatments and (2) the building of classiﬁers, with

machine learning techniques, which could be used to improve the diagnosis of patients

with cancer [6]. Bioinformatics has been proposing novel clustering methods that take

intrinsic characteristics of gene expression data into accounts, such as noise and high-

dimensionality, to improve the classiﬁcation [7,8].

Golub et al. [9] presented an effective method to identify a predictive gene subset

for cancer classiﬁcation. They used a neighborhood analysis in selecting a gene subset

that can distinguish between the two cancer types: AML and ALL, based on a sepa-

ration measure similar to t-statistic. In their experiments, the leukemia dataset contained

6,817 gene expression levels in 72 samples. The training samples were used to select a

gene subset that can distinguish between the two cancer types: AML and ALL. The

ﬁfty most predictive genes identiﬁed by the training samples were validated, to classify

the test samples.

Microarray technologies enable the measurement of the molecular signatures of cancer

cells. These data allow different types of analyses, such as (1) the identiﬁcation of

differentially expressed genes [5], which could indicate possible gene targets for more

detailed molecular studies or drug treatments and (2) the building of classiﬁers, with

machine learning techniques, which could be used to improve the diagnosis of patients

with cancer [6]. Bioinformatics has been proposing novel clustering methods that take

intrinsic characteristics of gene expression data into accounts, such as noise and high-

dimensionality, to improve the classiﬁcation [7,8].

Golub et al. [9] presented an effective method to identify a predictive gene subset

for cancer classiﬁcation. They used a neighborhood analysis in selecting a gene subset

that can distinguish between the two cancer types: AML and ALL, based on a sepa-

ration measure similar to t-statistic. In their experiments, the leukemia dataset contained

6,817 gene expression levels in 72 samples. The training samples were used to select a

gene subset that can distinguish between the two cancer types: AML and ALL. The

ﬁfty most predictive genes identiﬁed by the training samples were validated, to classify

the test samples.

Microarray technologies enable the measurement of the molecular signatures of cancer

cells. These data allow different types of analyses, such as (1) the identiﬁcation of

differentially expressed genes [5], which could indicate possible gene targets for more

detailed molecular studies or drug treatments and (2) the building of classiﬁers, with

machine learning techniques, which could be used to improve the diagnosis of patients

with cancer [6]. Bioinformatics has been proposing novel clustering methods that take

intrinsic characteristics of gene expression data into accounts, such as noise and high-

dimensionality, to improve the classiﬁcation [7,8].

Golub et al. [9] presented an effective method to identify a predictive gene subset

for cancer classiﬁcation. They used a neighborhood analysis in selecting a gene subset

that can distinguish between the two cancer types: AML and ALL, based on a sepa-

ration measure similar to t-statistic. In their experiments, the leukemia dataset contained

6,817 gene expression levels in 72 samples. The training samples were used to select a

gene subset that can distinguish between the two cancer types: AML and ALL. The

ﬁfty most predictive genes identiﬁed by the training samples were validated, to classify

the test samples.

Microarray technologies enable the measurement of the molecular signatures of cancer

cells. These data allow different types of analyses, such as (1) the identiﬁcation of

differentially expressed genes [5], which could indicate possible gene targets for more

detailed molecular studies or drug treatments and (2) the building of classiﬁers, with

machine learning techniques, which could be used to improve the diagnosis of patients

with cancer [6]. Bioinformatics has been proposing novel clustering methods that take

intrinsic characteristics of gene expression data into accounts, such as noise and high-

dimensionality, to improve the classiﬁcation [7,8].

Golub et al. [9] presented an effective method to identify a predictive gene subset

for cancer classiﬁcation. They used a neighborhood analysis in selecting a gene subset

that can distinguish between the two cancer types: AML and ALL, based on a sepa-

ration measure similar to t-statistic. In their experiments, the leukemia dataset contained

6,817 gene expression levels in 72 samples. The training samples were used to select a

gene subset that can distinguish between the two cancer types: AML and ALL. The

ﬁfty most predictive genes identiﬁed by the training samples were validated, to classify

the test samples.

Microarray technologies enable the measurement of the molecular signatures of cancer cells. These data allow different types of analyses, such as the identiﬁcation of differentially expressed genes, which could indicate possible gene targets for more detailed molecular studies or drug treatments and the building of classiﬁers, with machine learning techniques, which could be used to improve the diagnosis of patients

with cancer. Bioinformatics has been proposing novel clustering methods that take intrinsic characteristics of gene expression data into account, such as noise and high dimensionality, to improve the classiﬁcation.

Presented an effective method to identify a predictive gene subset for cancer classiﬁcation. They used a neighborhood analysis in selecting a gene subset that can distinguish between the two cancer types: AML and ALL, based on a separation measure similar to the t-statistic. In their experiments, the leukemia dataset contained 6,817 gene expression levels in 72 samples. The training samples were used to select a gene subset that can distinguish between the two cancer types: AML and ALL.

The ﬁfty most predictive genes identiﬁed by the training samples were validated, to classify the test samples.

* **Model architecture**

**K-Means Clustering Model:**

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on. It allows us to cluster the data into different groups and is a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training. It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters. The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

* Determines the best value for K center points or centroids by an iterative process.
* Assigns each data point to its closest k-center. Those data points which are near toe particular k-center, create a cluster.

The working of the K-Means algorithm is explained in the below steps:

* Step 1: Select the number K to decide the number of clusters.
* Step 2: Select random K points or centroids. (It can be other from the input dataset).
* Step 3: Assign each data point to their closest centroid, which will form the predefined K clusters.
* Step 4: Calculate the variance and place a new centroid of each cluster.
* Step 5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.
* Step 6: If any reassignment occurs, then go to step 4 else go to FINISH.
* Step 7: The model is ready.

**Logistic Regression Model:**

Logistic regression has become an important tool in the discipline of [machine learning](https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML). It allows algorithms used in machine learning applications to classify incoming data based on historical data. As additional relevant data comes in, the algorithms get better at predicting classifications within data sets.

It can also play a role in [data preparation](https://www.techtarget.com/searchbusinessanalytics/definition/data-preparation) activities by allowing data sets to be put into specifically predefined buckets during the extract, transform, load ([ETL](https://www.techtarget.com/searchdatamanagement/definition/Extract-Load-Transform-ELT)) Process in order to stage the information for analysis.

Logistic regression can also estimate the probabilities of events, including determining a relationship between features and the probabilities of outcomes. That is, it can be used for classification by creating a model that correlates the hours studied with the likelihood the student passes or fails. On the flip side, the same model could be used for predicting whether a particular student will pass or fail when the number of hours studied is provided as a feature and the variable for the response has two values: pass and fail.

**Support Vector Machine**

“Support Vector Machine” (SVM) is a supervised machine learning algorithm that can be used for both classification and regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

### Random Forest

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

Steps involved in random forest algorithm:

Step 1: In Random forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

### XG Boost

'Boosting' Algorithms are one of the most widely used algorithms and they basically grant power to machine learning models to improve their accuracy of prediction. Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made. 'It can be refereed to as a family of algorithms which converts weak learners (weak prediction models) to strong learners(strong prediction models)

XGBoost is an implementation of Gradient Boosted decision trees. In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

**With Grid Search**

One method is to try out different values and then pick the value that gives the best score. This technique is known as a grid search. If we had to select the values for two or more parameters, we would evaluate all combinations of the sets of values thus forming a grid of values.

**With PCA**

To identify features to be used for supervised learning.

We use PCA as part of preprocessing, the algorithm is applied to:

1. Reduce the number of dimensions in the training dataset.
2. De-noise the data. Because PCA is computed by finding the components which explain the greatest amount of variance, it captures the signal in the data and omits the noise.

**Decision Tree**

In a decision tree, for predicting the class of the given dataset, the algorithm starts from tree’s root node. This algorithm compares the values of the root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.

For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm:

* Step-1: Begin the tree with the root node, says S, which contains the complete dataset.
* Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).
* Step-3: Divide the S into subsets that contains possible values for the best attributes.
* Step-4: Generate the decision tree node, which contains the best attribute.
* Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

KNN Model

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset.

The K-NN working can be explained on the basis of the below algorithm:

* Step-1: Select the number K of the neighbors
* Step-2: Calculate the Euclidean distance of K number of neighbors
* Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.
* Step-4: Among these k neighbors, count the number of the data points in each category.
* Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.
* Step-6: Our model is ready.

**Neural Network using Keras**

What is Keras?

* Keras is a high-level neural network API that is written in Python.
* It is capable of running on top of Tensorflow, CNTK, or Theano.
* Keras can be used as a deep-learning library. Support Convolutional and Recurrent Neural Networks
* Prototyping with Keras is fast and easy
* Runs seamlessly on CPU and GPU

**LDA Model**

The goal of LDA is to project the features in higher dimensional space onto a lower-dimensional space in order to avoid the curse of dimensionality and also reduce resources and dimensional costs. LDA is a supervised classification technique that is considered a part of crafting competitive machine learning models. Linear Discriminant analysis is used as a dimensionality reduction technique in machine learning, using which we can easily transform a 2-D and 3-D graph into a 1-dimensional plane.

Let's consider an example where we have two classes in a 2-D plane having an X-Y axis, and we need to classify them efficiently. As we have already seen in the above example that LDA enables us to draw a straight line that can completely separate the two classes of data points. Here, LDA uses an X-Y axis to create a new axis by separating them using a straight line and projecting data onto a new axis.

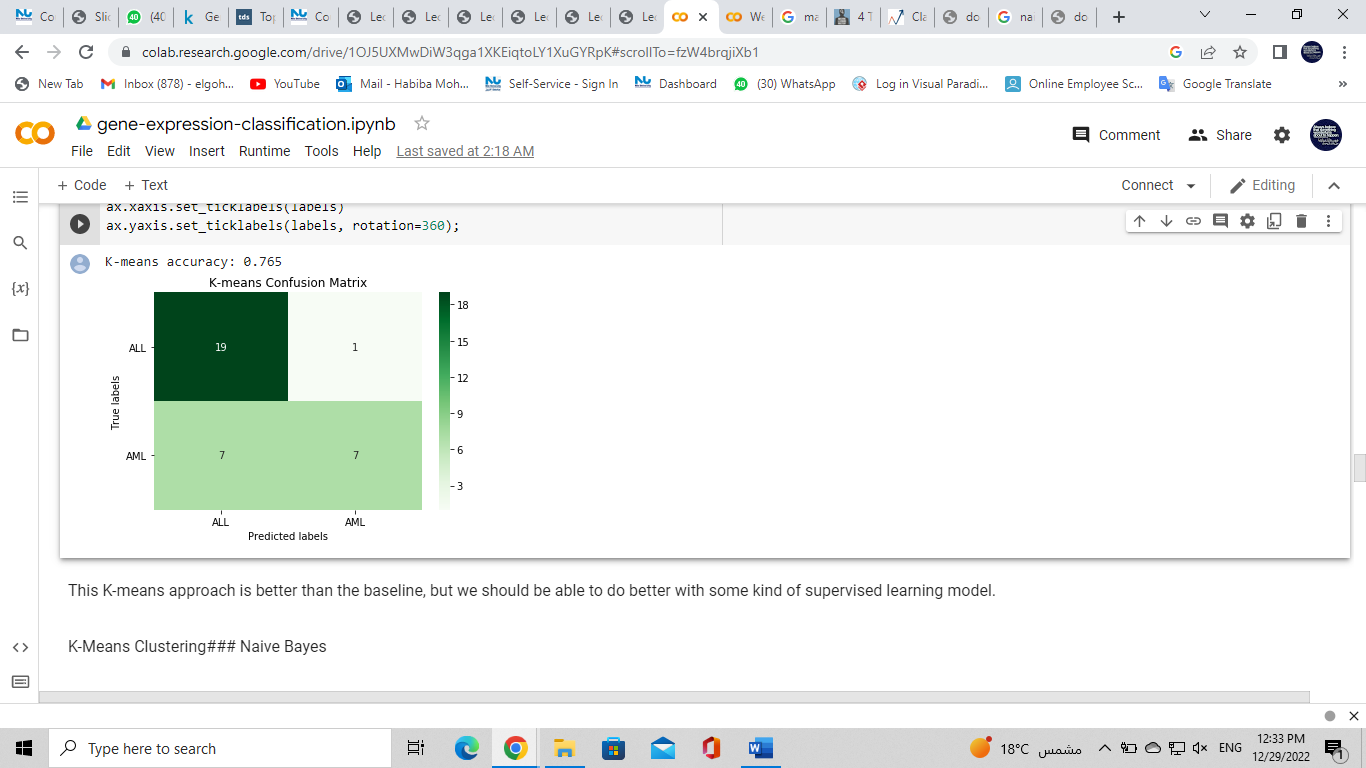
Hence, we can maximize the separation between these classes and reduce the 2-D plane into 1-D.

* It maximizes the distance between the means of two classes.
* It minimizes the variance within the individual class.

Using the above two conditions, LDA generates a new axis in such a way that it can maximize the distance between the means of the two classes and minimizes the variation within each class.

In other words, we can say that the new axis will increase the separation between the data points of the two classes and plot them onto the new axis.

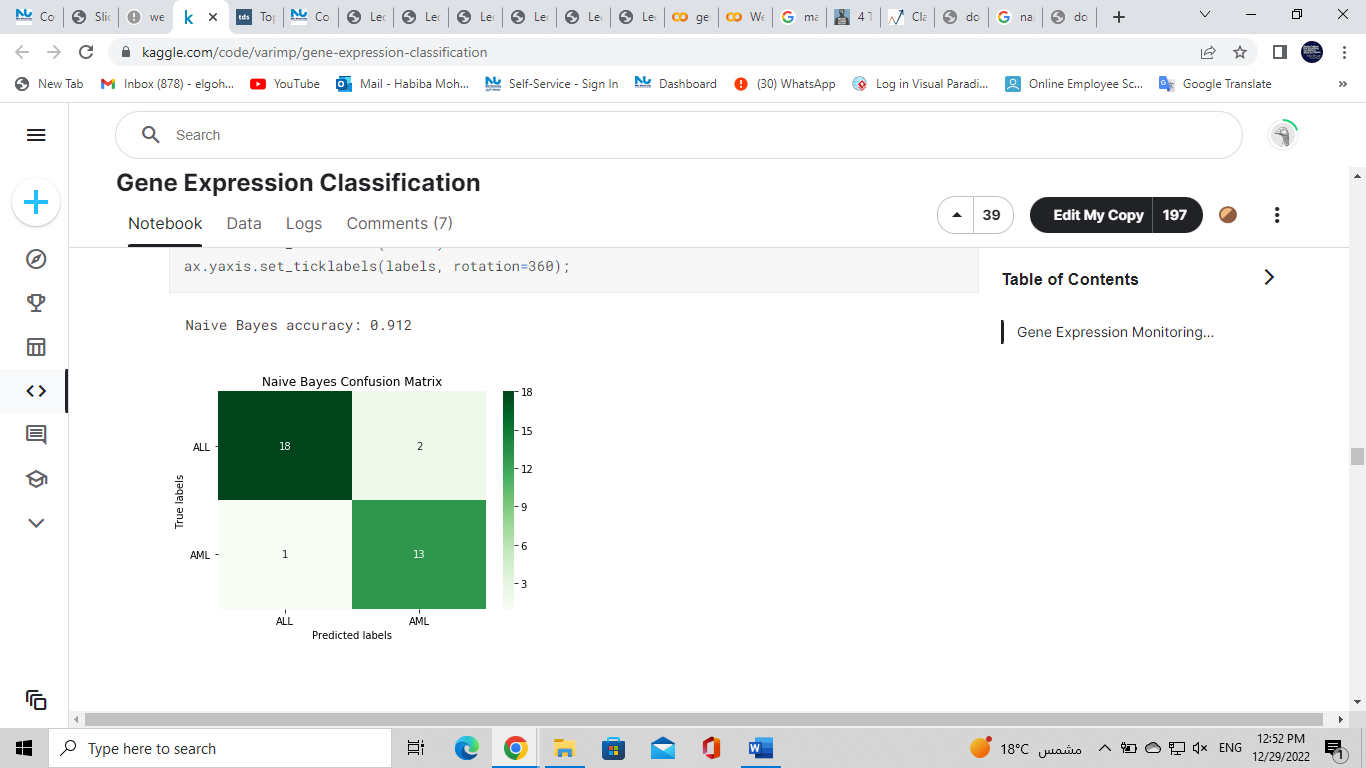
* **Evaluation results and strategies.**

We used 10 models to classify leukemia patients into one of two classes, unsupervised clustering approach and supervised clustering approach.

1. **K-means**

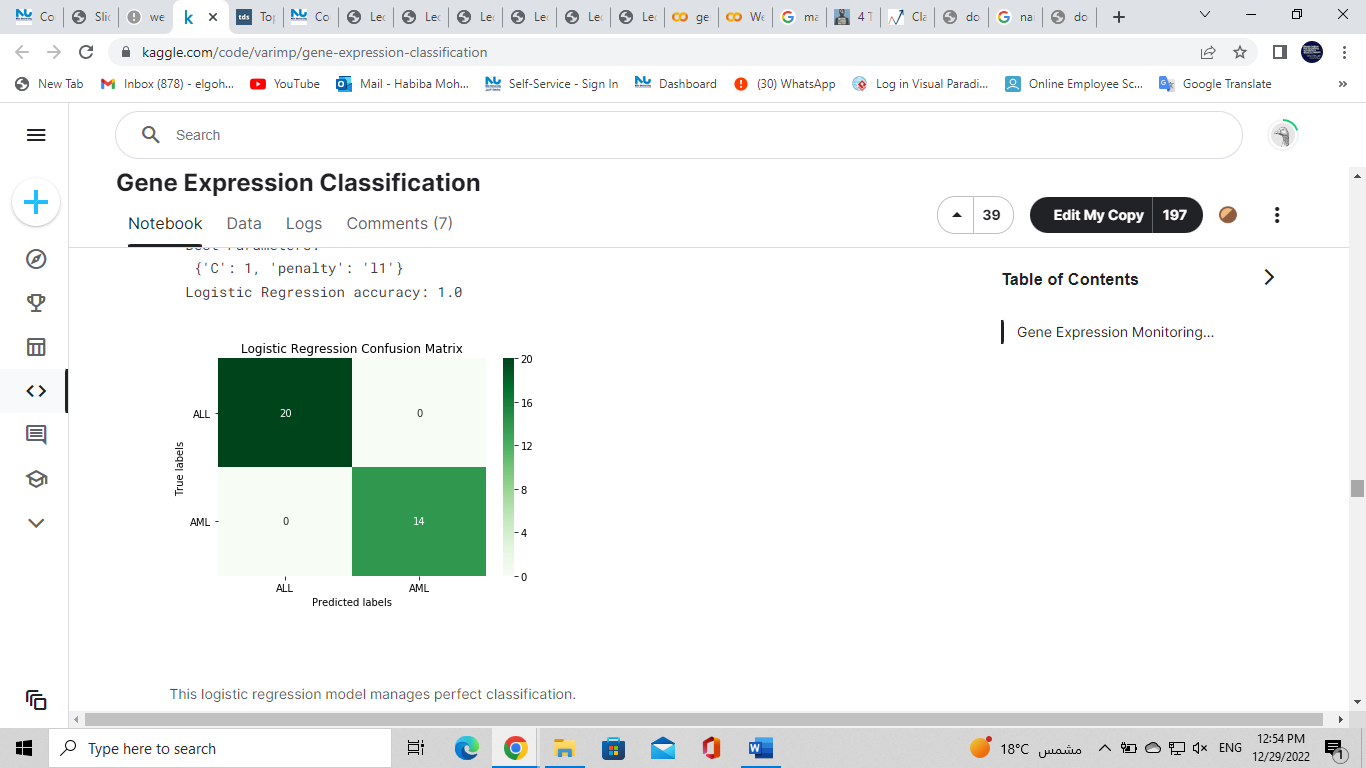
This K-means approach is better than the baseline, but we should be able to do better with some kind of supervised learning model.

Accuracy: 0.765

1. **Naïve bayes**

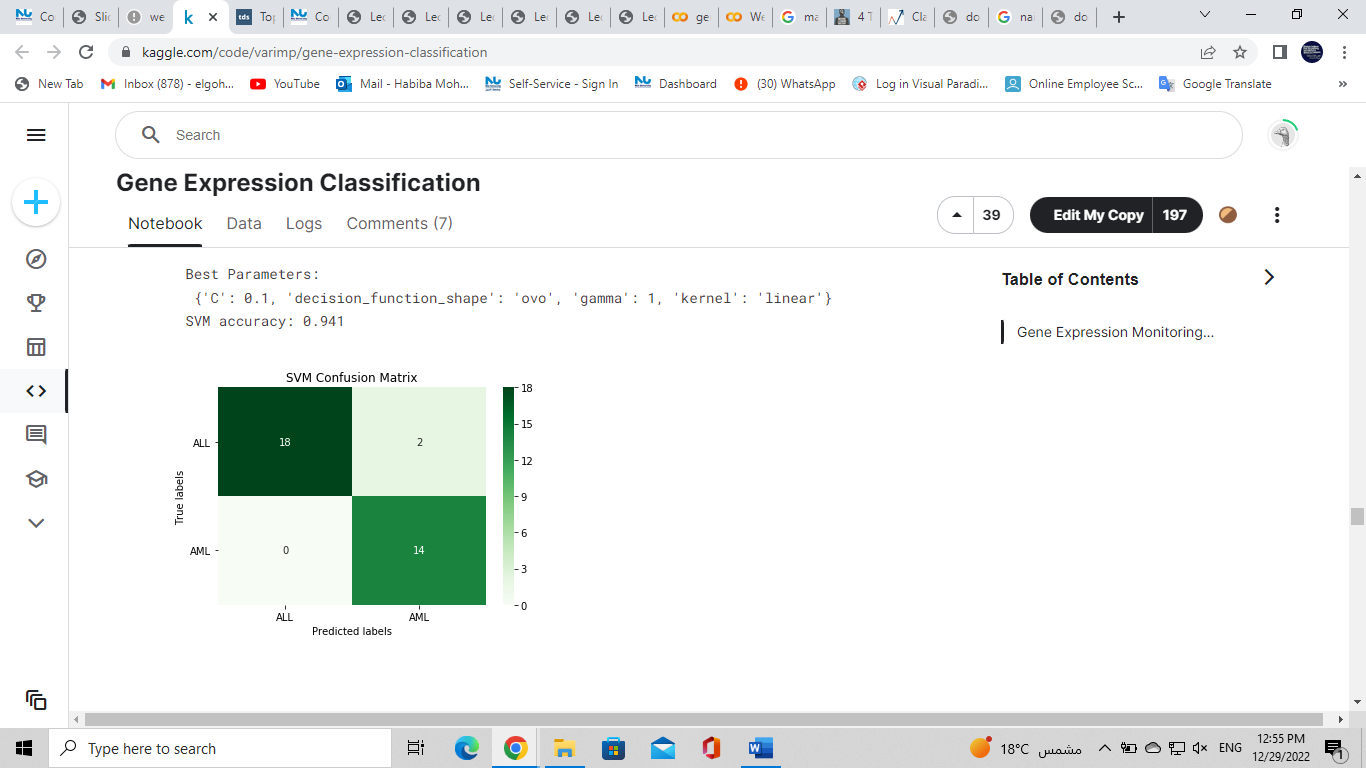
The naive bayes model is pretty good, just three incorrect classifications.

Accuracy= 0.912

**3** . **Logistic Regression**

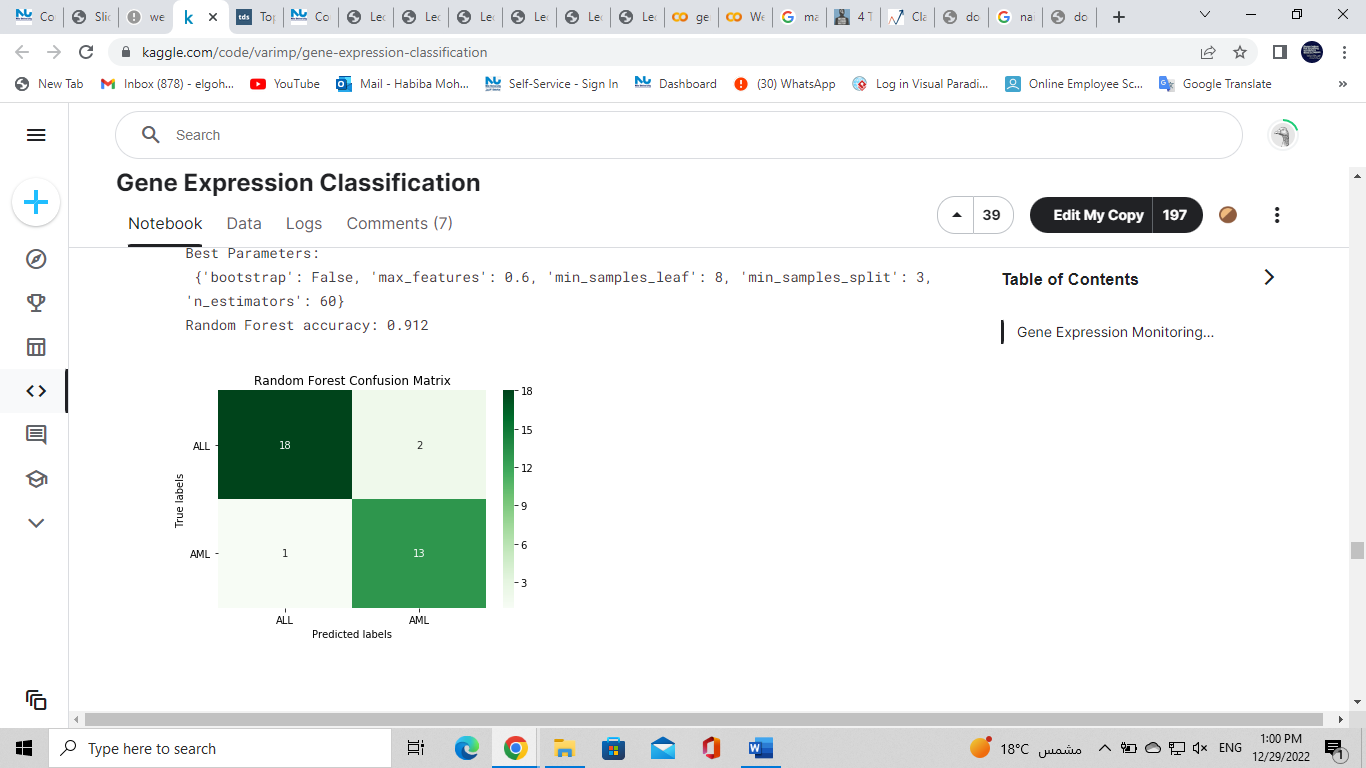
manages perfect classification.

Accuracy= 1.0



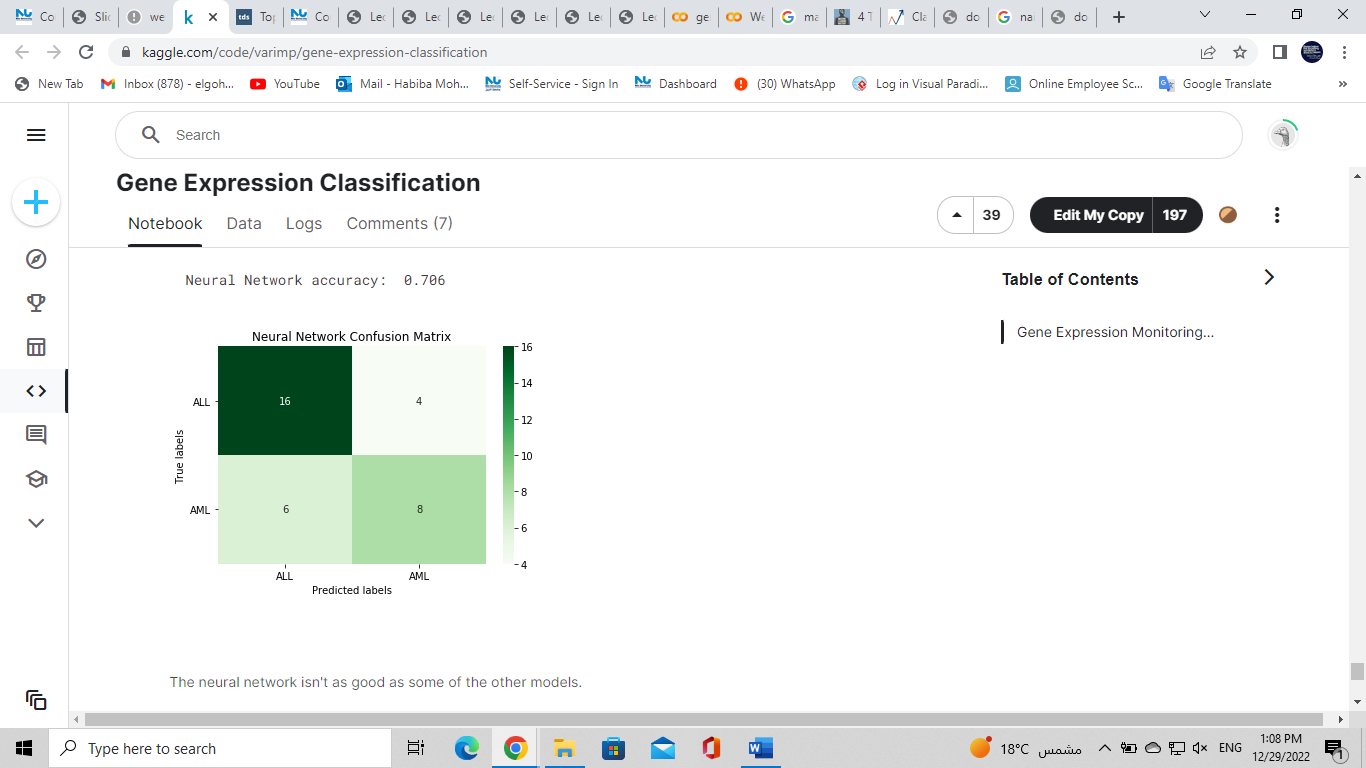
1. **Support vector machine:**

Accuracy= 0.941



1. **Random forest**

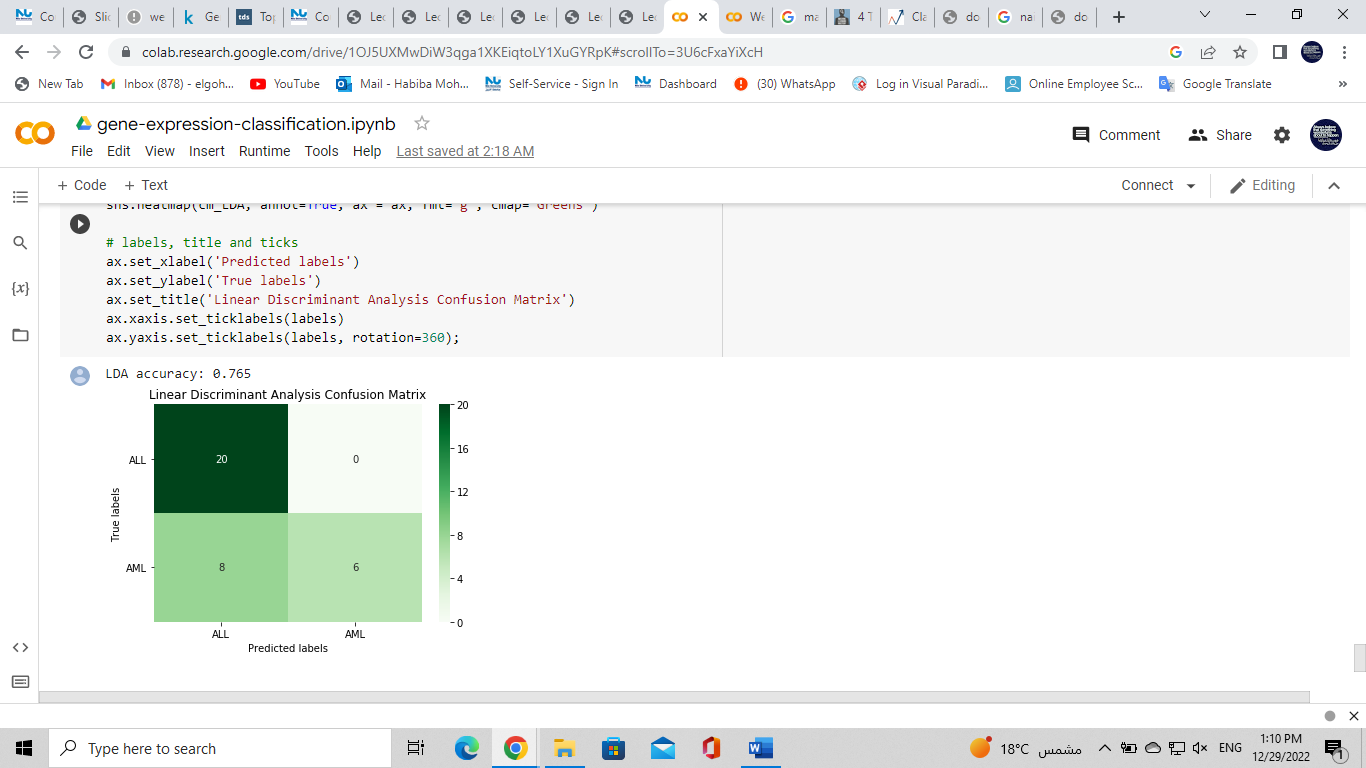
0.912

1. **Neural network**

Accuracy = 0.706

1. **LDA Model**

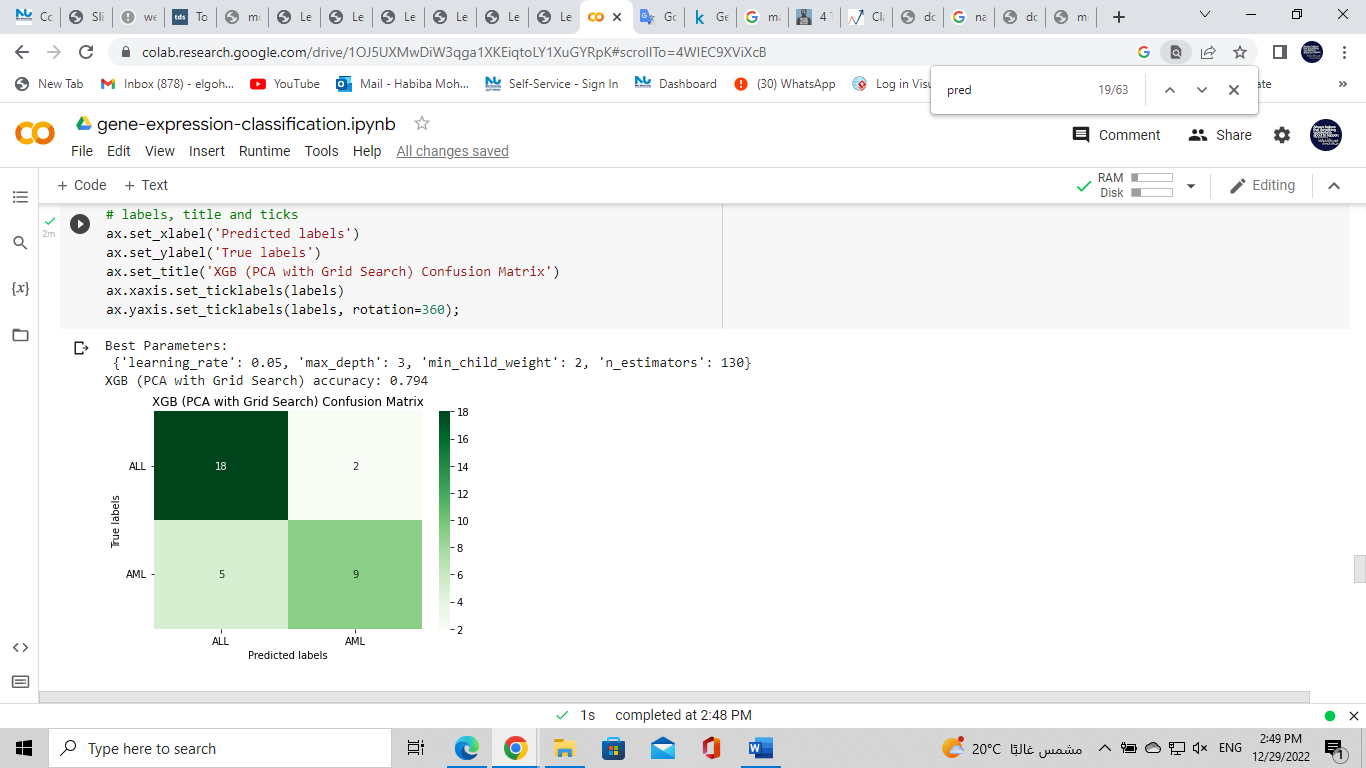
**Accuracy= 0.765**



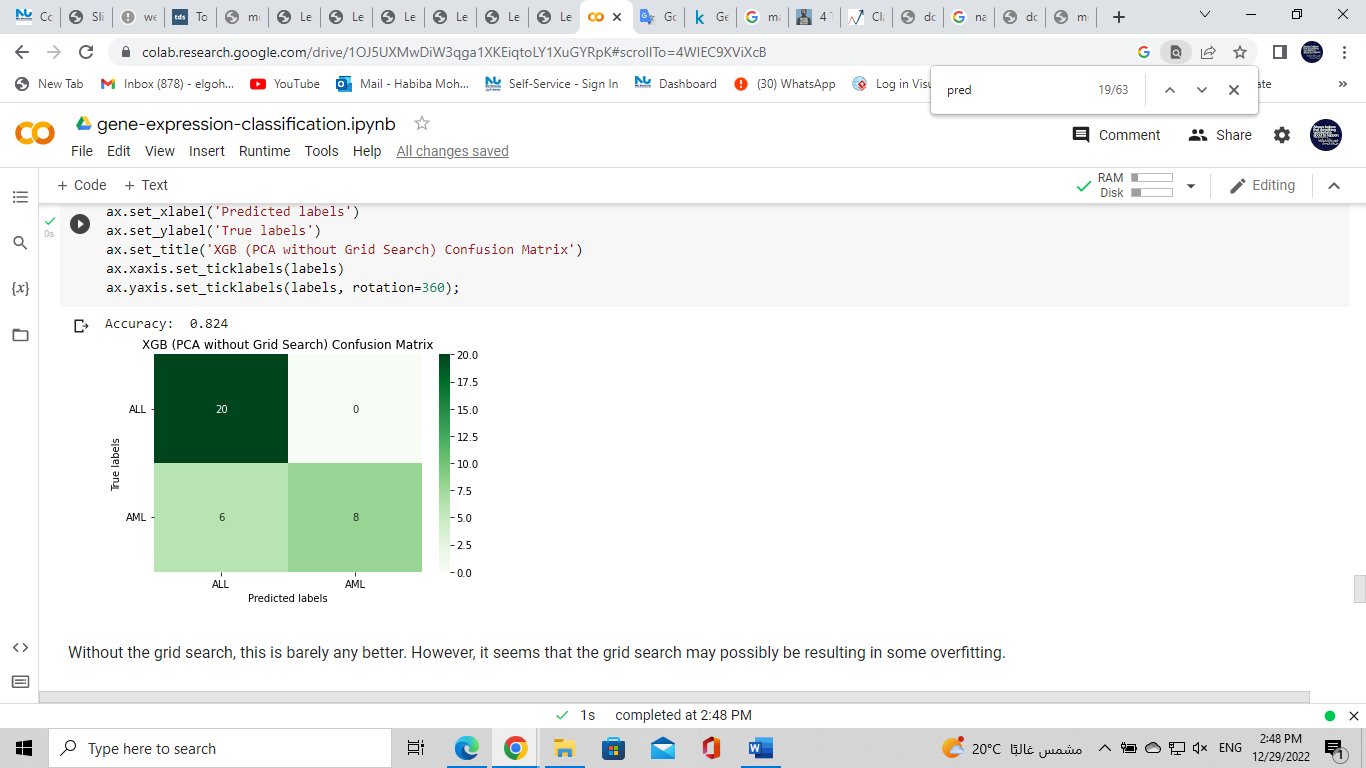
### XG Boost

We experiment with three alternative versions, PCA with grid search, PCA without grid search and also the original data without either PCA or grid search.

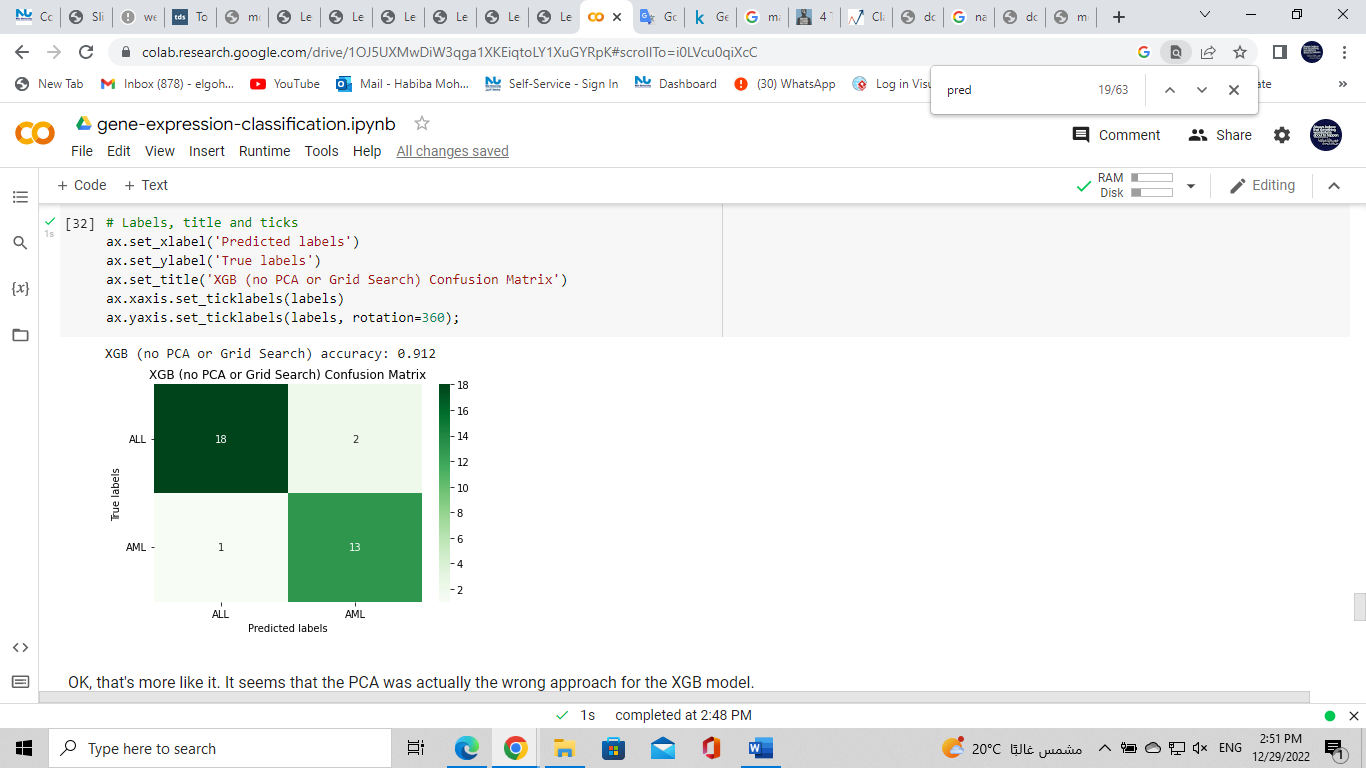
1. **XGB with** PCA and grid search isn't particularly good.



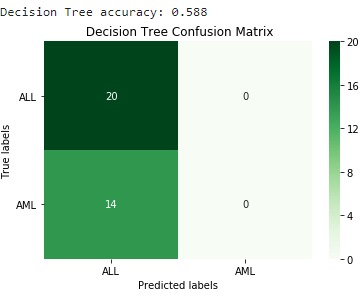
1. **XGB — PCA with no Grid Search** , this is barely any better



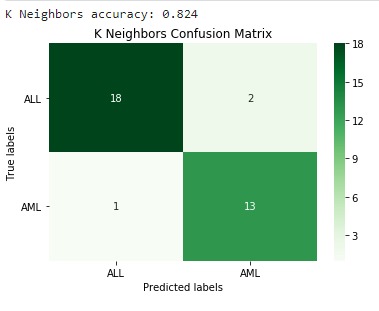
**3. XGB — no PCA or Grid Search**



1. **Decision tree:**

****

1. **K neighbours:**

****

**Evaluation strategies:**

Accuracy is the conventional method of evaluating classification models, accuracy is defined as the proportion of correctly classified examples over the whole set of examples, and accuracy is often used in conjunction with other methods. One way to check whether you can use accuracy as a metric is to construct a confusion matrix , which is an error matrix.

Best model used is Logistic regression as it has the highest accuracy =1, then Support vector machine= 0.94.

* **GitHub link**