*NETWORK BASED DATA ANALYSIS REPORT*

*Andrea Policano 24/07/2024*

*“Analysis of clustering methodologies on Colorectal Cancer samples”*

1. ***ABSTRACT***

At the moment many are the methodologies applied to study and remodel data while also trying to maximize the computational efficiency and the final results obtained. Of these techniques, many are used during health-related analysis, either to discover unknown/hidden pathways or patterns, reduce the overall complexity of the data in use, personalize and upgrade prediction models or to simply evaluate the quality and consistency of the information and results produced. Other advantages can be seen in the combination of multiple methodologies for further improving the quality of the solutions proposed. In our case the objective will be to compare and assess how different clustering approaches infer while using a disease-related dataset, connected to the Colorectal Cancer, and how will they be able to replicate the effective subdivision of the samples setting them side by side with the grouping that can be obtained by retrieving the metadata of said dataset.

1. ***INTRODUCTION***
   1. ***COLORECTAL CANCER (CRC)***

**Immagine che contiene testo, schermata

Descrizione generata automaticamente**CRC is a type of cancer characterized by the abnormal growth of cells in the first and longest part of the large intestine, called “Colon”. Even though it may become present at any age, it is frequently present in age-advanced individuals, usually starting from the formation of small cell lumps, called “polyps” that can be easily identified inside the colon, they do not cause symptoms neither are directly dangerous, although they possess the potential to become cancerous elements, their removal is also associated with better recovery against the cancer itself. Some of the symptoms that may arise and that can be related to CRC are:

Figure 1, Colon section of the Large Intestine

* **Change in bowel habits**: frequent diarrhoea or constipation
* **Rectal bleeding**
* **Belly discomfort**: characterized by cramps, gas or general pain
* **Weakness**
* **Loss of blood from the rectum**
* **Loss of weight**
* **Bowel un-emptiness**: feeling that the bowel is still empty after peristaltic movements

As stated, one of the main factors associated to CRC is the old age of the patient, other risk factors that go hand in hand with it are:

* **RACE**: the disease seems to have a link to the race of the individual, especially for Black people in the US
* **Family and personal history**
* **Inflammatory bowel diseases**: such as Crohn’s disease or ulcerative colitis
* **Inherited disease**
* **Low-fiber, high-fat diet**: typical Western diet, also results were seen to be linked with processed and red meat
* **Absence of exercise**
* **Obesity**
* **Diabetes**
* **Smoking**
* **Alcohol consumption**
  1. ***GSE52060***

The dataset here presented is the one used in the beforementioned study, it is derived and produced by 46 samples coming from patients presenting Colorectal Cancer (CRC), from them one sample coming from the neoplastic tissue and one coming from the normal mucosal tissue were taken, later on these two subdivisions will become the main factors for methodology comparisons and analysis.

The GSE file present on GEO (Gene Expression Omnibus) was fetched and loaded on the virtual environment of Rstudio, from it pheno-related data was extracted and a metadata dataframe was built, mainly used to understand additional information regarding the samples, how they are grouped and possibly find some potential targets to use as factors for later stages in analysis. From that we can see that we have 23 samples of Normal mucosa and 23 of Neoplastic tissue, for a grand total of 46 analysable samples.

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

Figure 2, Metadata dataframe obtained from the GSE52060 file

* 1. ***CLUSTERING TECHNIQUES INVOLVED***

Throughout the study different clustering techniques will be applied in order to understand how efficient they are in a sample classification task, considering the given dataset and how they will fare individually. Those which will be tested are: PCA, K-means and Hierarchical clustering, Random Forest, Linear Discriminant Analysis, LASSO and RSCUDO. Each of these having their own methodology and way of working, which will “triumph” over the others given the situation?

1. ***METHODS***
   1. ***EXPRESSION MATRIX RETRIEVAL***

Immagine che contiene testo, schermata, linea, tipografia

Descrizione generata automaticamenteAfter loading the GSE file and retrieving the metadata dataframe, the next step in the project was to obtain the expression matrix of the data present in the file. That was done swiftly by few command lines. After the acquisition, results of the expression matrix were box plotted, picturing a before-and-after situation in the context of a log2 normalization in which an improved organization of the data can be seen.

Figure 3, Boxplot of pre-normalized GSE52060 expression data

Immagine che contiene testo, comb

Descrizione generata automaticamente

Figure 4, Boxplot of pre-normalized GSE52060 expression data

* 1. ***PRINCIPAL COMPONENT ANALYSIS (PCA)***

PCA is a machine learning method focused on the dimensionality reduction in order to simplify a large dataset into a smaller one, maintaining all the important information, pattern and trends. This methodology of course still pays the price when applied, the reduction itself will cost us accuracy but in the project’s context it is highly profitable for the identification of homogeneous subgroup of genes with similar expression profiles or samples which present akin trends.

PCA was performed through the use the *prcomp* command together with the execution of a t-test on the dataframe used in order to maximize gene prioritization and feature selection, following these processes both a summary and a screeplot were produced together with other graphs produced using the *autoplot* function,the colours were assigned by taking in consideration the column 8 of the *metadata\_df* variable “source\_name\_ch1” which showed the subdivision of the two tissues recovered.

Immagine che contiene testo, schermata, diagramma, linea

Descrizione generata automaticamente

Figure 5, PCA of metadata dataframe using as reference the two tissue types, in this case the “skeleton” of the PCA can be seen, together with the absence of obvious clustering or grouping of the samples due to the lack of characterization using colours

* 1. ***HIERARCHICAL AND K-MEANS CUSTERING***

Following the creation of the PCA plot, the next step of the analysis revolved around the use of clustering methods for further information retrieval and processing.

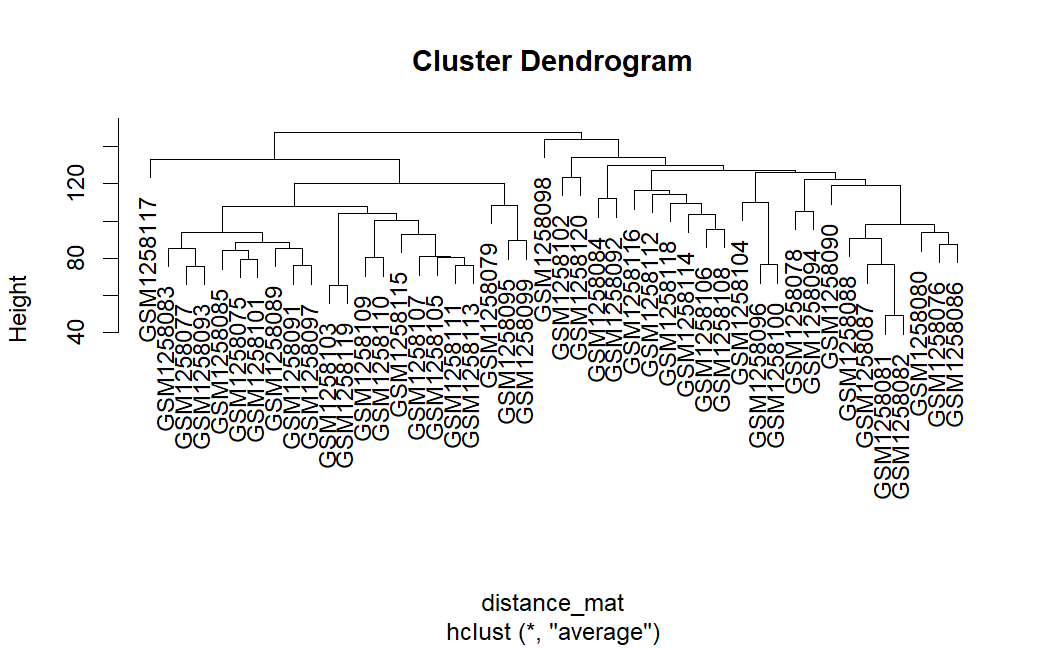


Figure 6, Hierarchical clustering raw plot, showing the respective distance measures for each sample and hints on their relationships

* + 1. ***CLUSTERING***

It is an unsupervised machine learning methodology, in which no class values hinting at *a priori* grouping of the data are given. It focuses on finding similarity groups in a data set and it is a daily-life technique that is commonly put in practice, some of its day-to-day applications are for example: grouping clothes, objects, videogames, tools, documents or even target possible marketing subjects.

It is characterized by an algorithm in which a distance is contained, the quality of a clustering process is measured by taking in consideration some the data, the distance function and in general the algorithm itself, as in this case we ought to consider a trade-off between accuracy/quality and computational feasibility.

* + 1. ***K-MEANS CLUSTERING***

K-means is a partitional clustering methodology used to assign data points to one of the K clusters depending on the distance of the points themselves. The algorithm starts by randomly assigning cluster centroids in the space and grouping the datapoints to them, while taking in consideration the distance from the centre of the cluster. After this point new centroids are selected and the process is repeated iteratively, until the change in the Sum of Squared Errors (SSE) goes below a certain threshold, or the iterations are completed.

It is highly effective and time-convenient, but it also presents its downsides, such as the need of an optimal K to choose in much more complicated case, when few clusters are not suitable for the study, or the need to define the *mean* and its sensitivity to outliers but it is highly effective in many situations.

In the project context the “*K-means*” algorithm was used in order to distinguish the samples using the tissue types to which they belonged, the K-value used was 2 as the maximum number of tissues. This step produced a plot similar to that of the PCA one, with the presence of the label of the samples and the same colouring used in the phase before.

* + 1. ***HIERARCHICAL CLUSTERING***

As for the one before is a clustering algorithm, with the only difference being the use, in the Hierarchical, of a distance matrix, in which the distance measure can be defined by different methods such as: the Euclidian, Manhattan or the Minkowski distance, with the last being used only in specific cases. Other than this we ought to take in consideration also the production of a different end result with respect to the *k-mean* one, by applying this methodology we obtain a tree called *dendrogram* with the samples taking the part of the leaves.

Hierarchical clustering can function both as a bottom-up or a top-down approach, using either a single or complete-linkage approach. In the project this technique was used to produce different dendrograms to see how the samples are related, taking into account, as in the steps before, their tissue subdivisions. In this case the main command used in this phase are: *hclust* for the tree production with the methodology “*ave*”; *cutree* for squads/groups production using the metadata; *rect*.*hclust* for dendrogram production with a border identifying the two groups and finally the production of the dendrogram of the samples with the branches colours related to the tissue types.

Immagine che contiene testo, diagramma, Piano

Descrizione generata automaticamente

Figure 7, One of the dendrogram produced by applying Hierarchical Clustering, here is one the three dendrogram produced and represent the starting point for the others production

* 1. ***HEATMAP***

A heatmap is a 2-dimensional data visualization technique in which the data are represented through the use of a gradient of colours, usually using a darker one for greater quantity of that value, the main aim of this procedure is that of finding how the samples are correlated between each other and to exploit these relationships in future steps.

As for other common methodologies it can be used in many different fields, going from criminology to gene-expression analysis, with lots of different type of it depending on the study type in which the methodology is applied.

The first thing done before the heatmap production was creating a distance matrix with the data coming from the previous step, along with the setting of a colour palette that will be needed for the final plot. The *pheatmap* function was used for the creation of the heatmap using as main data the distance matrix, while incorporating the already available *distance\_mat* variable obtained in the Hierarchical Clustering step, and the palette of colours.

* 1. ***RANDOM FORESTS***

Even though they are not directly correlated to clustering algorithms, they still play a role in classification and regression operations in machine learning models. Making use of the ensemble mechanism “bagging” they are based on the production of many decision trees which will later be merged in order to obtain a final model able to have a much more accurate output value, increasing the quality of the learning model. An additional feature that separates this type of algorithm from the rest is the use of labelled data, thus showing how its origin is related to the supervised learning typology.

Its application in our project starts with the creation of a new dataset that will be specifically used only for this phase of the project, in concomitance the conditions for the dataset’s related samples were obtained and set from the metadata\_df obtained in the first part of the analysis, the conditions are the same that we used for the other cases: Neoplastic and Normal mucosal tissues. Following along the model was produced using the *randomForest* function, later on, graph related to the model were also produced, focusing on the most variable relevance(*fig.6*) and the main probes useful for the model quality(*fig.7*).

Immagine che contiene testo, schermata, diagramma, linea

Descrizione generata automaticamente

Figure 8, Plot showing the relevance of the first samples of the used dataframe in the quality of the model, meaning that the only the starting samples are relevant for the model

Immagine che contiene testo, schermata, linea, numero

Descrizione generata automaticamente

Figure 9, Plot showing which are the probe having the highest relevance/importance in the model created, with the addition of the visualization of how much the loss of quality will affect the model if that probe were to be modified

* 1. ***LINEAR DISCRIMINANT ANALYSIS (LDA)***

LDA is a supervised machine learning approach that aims to solve multi-class classification problems by separating them through data dimensionality reduction. It is especially useful to optimize machine learning models.

This technique is able to make predictions by using Bayes theorem and calculate the probability of a particular data set to belong to an output, while also modelling the data distribution. It focuses on maximizing the between-class distance and minimizing the within-class one, in addition LDA works by identifying a linear combination of features that can be used to separate two or more classes of objects. It is done by projecting the data into a one-dimensional graph for easier classification, proving a high level of versatility also for multi-class data.

In order to evaluate the classifier obtained through the LDA the ROC and CARET methodologies were used. At the beginning of this step the creation of factors *"tissue type: adjacent normal mucosa (N)"* and *"tissue type: neoplastic tissue (T)"* was completed, after that t-tests were conducted on the expression dataframe using the function *rowttests* while also accounting for the new factors, extraction of probes having a p-value lower than 0.05 followed and a new dataframe was built, additionally a column was inserted in the df, called “*AFFECTED*” hinting whether or not the sample is derived from the neoplastic tissue.

Immagine che contiene testo, diagramma, schermata, linea

Descrizione generata automaticamente

*Figure 10, LDA derived plot showing the separation between the samples clusters, here are visible some elements that may result in outliers*

Following this path the model went under training using the function *lda*, taking as input parameters:

* *AFFECTED* (as Function): needed for understanding which are the variables/samples that we want to guess as having neoplastic tissue against the normal mucosal one
* The dataframe created before that has built-in the factors of interest
* The prior probability of each class
* The subset of elements that we want to train

After these steps were completed, the mod.values were predicted using the function *predict* coming from the , taking as input the model built and the dataframe’s training subset.

* + 1. ***CARET***

After completing the first part of the LDA phase, the Caret methodology was implemented. It provides a wide range of functions focused on data preparation, modelling and evaluation. With the objective to streamline the evaluation and building process of predictive models, the package used also includes functions related to data preparation, modelling and evaluation.

CARET mainly works by working on Hold-out specific samples and fitting the model on the remaining ones, in the end the average performance is calculated across all the hold-out predictions and the determinations of the optimal parameters set is given. When starting with the application of CARET, a control group and a metric needed to be defined:

1. **Control** 🡪 built by the *trainControl* function using as input:
   1. *Method*: “*cv*”, setting the resampling method to cross-validation
   2. *Number*: “*10*”, indicating the 10-fold cross-validation
   3. *Repeats*: “*NA*”, iteration of the methods
2. **Metric** 🡪 set as “*Accuracy*”

Later training of the two models was conducted by using the *train* function using the *lda* and *rf* methodologies and in both cases considering also the metric established. Results were obtained using also the *resamples* function and *ggplot* for the representations.

* + 1. ***ROC CURVE***

The ROC Curve is a probability graph able to show the classification model performance considering all the classification thresholds, potting two parameters: TPR (True Positive Rate) vs FPR (False Positive Rate), at different thresholds meanwhile it also tries to separate the “*signal”* from the potential “*noise”* detected.

It is used mainly to evaluate the performance of a binary classifier considering the two metrics already cited:

* **TPR** 🡪 Sensitivity, proportion of correctly identified positives that are actual ones
* **FPR** 🡪 Measuring the proportion of actual negatives incorrectly identified by the model

Results in this part of the experiment by simply redoing a prediction of the model, using in this case the test subset, and plotting thir results using the *plot.roc* function. In addition to that the value of AUC (Area Under the Curve) was calculated.

* 1. ***LASSO/RIDGE REGRESSION***

These two are very famous algorithms heavily used in the field of machine learning, both of them are used to reduce errors deriving from a linear regression model, with each one presenting a different key feature that characterize it.

* RIDGE REGRESSION 🡪 introduces a regularized (penalty) term (λ) in the cost function in order to prevent overfitting, it can reduce all the coefficients by a small amount
* LASSO REGRESSION 🡪 also in this case the penalty term is present but in this case is regularization L1 instead of L2 like in the RIDGE case. L1 represent the sum of the absolute values of the coefficients then multiplied by a constant λ, in this case the main objective is to reduce the features.

Immagine che contiene testo, Diagramma, linea, diagramma

Descrizione generata automaticamente

Figure 11, Plot showing the trend of the quality of the model depending on the chosen regularization parameter

During the study LASSO was the used algorithm instead of the RIDGE due to problems and anomalies related to the data sets. In this case we reutilize the function *train* already implemented in the CARET steps changing some of the input parameters:

1. **Function** 🡪 *AFFECTED*, recalling the LDA methods
2. **Data** 🡪 the model matrix deriving from LDA
3. **Method** 🡪 *glmnet*
4. **Family** 🡪*binomial*
5. **Alpha** 🡪 1, indicating the penalty for LASSO
6. **Control** 🡪 *control*, recalling the CARET section
7. **Metric** 🡪 the metric already produced in the previous part of the project

In the end, plots regarding the relationship between coefficient, binomial deviance and log lambda were produced and the model was compared to the ones obtained from RF and LDA.

* 1. ***RSCUDO***

After the implementation of LASSO, the next algorithm taken into consideration and used in the project was RSCUDO. It implements decision trees in order to identify robust subgroups through iterative clustering, it is mainly used in the biological context, focusing specifically on transcriptomics and genomic data.

In this project the whole phase starts with the factorization (all over again) of the conditions of the tissue and creation of both a train and test group, the following step led to the training through the specific *scudoTrain* function, signatures check was the successive event in the phase, subsequently the creation of different networks took place. The criteria for the differentiation of the cluster created was the type of subdata used in the conditions and a classification procedure carried out only in the final graph, based on the factorized *in\_training* data.

Immagine che contiene schermata, Elementi grafici, design

Descrizione generata automaticamente

Figure 12, Example of network built while using the RSCUDO package, here can be seen the main two clusters and some elements "on the line" between them

* 1. ***PATHFINDR***

With the implementation on the starting dataset of all the wanted clustering/classification algorithms, pathfindR was applied in order to study the enrichment results of the used genes in the dataset. Before doing any type of operation in this scenario a filtered list (p-value < 0.01) of genes was extracted from the df, which will later be used in the manual enrichment carried out using the software enrichR available online (*see 3.10.*)

Once done with the extraction, the process for pathfinder application begins. The starting point is represented by the developing of a model matrix based on the conditions of the tissue, followed by a “fit model” operation and production of a contrast matrix, lateron an analysis on said matrix was carried out using *limma,* from which we could obtain the list of probes of interest, further filtered by p-value (<0.01)*.*

Until now all the operations that were enacted had a role in the formation of the final dataframe that would later be used for the effective run of pathfinder, still before that a merge between the results obtained until now and another df, created through the use of both probeID and their respective gene symbols, associated through the use of *ensemble* and *biomart.*

Subsequently to the merging, different databases were opted to run the algorithm, in the end 3 of them were chosen: KEGG, GO-BP, Reactome. With everything done and ready to go pathfinder was applied through the *run\_pathfindR* function for each of the selected databases, together with the upcoming creation of a plot for each case.

* 1. ***ENRICHMENT***

After obtaining the df having the ILLUMINA probes, a new empty dataframe was built. From previous analysis the logFC and the p-value of each probe were extracted and implemented in the new structure, in addition a column having the gene symbol of the probes was created through the use of the “*illuminaHumanv3.db*” and “*Annotation.dbi*” libraries and the *mapIDs* function, able map the probe id to the respective gene symbol using the given database, the inputs were as such:

* **Database** 🡪 “*illuminaHumanv3.db”*
* **Keys** 🡪 “probe\_IDs” identify the list of elements that you want to convert, in this case the column having the probes ID
* **Columns** 🡪 “*SYMBOL*”, which type of column to take in consideration for the retrieval from the db
* **Keytype** 🡪 “*PROBID*”, indicating what is the object that you want to be converted

Concluding the mapping phase, a df having as columns both the probe and gene symbol IDs were obtained in the end, leading to the merging of the newly obtained element with a modified, t-tested df retrieved from the starting data and metadata. Concluding the merging between the two we will have a new structure in which the IDs will be present together with statistics such as p-value, logFC and dm.

From that point the objective is to retrieve the gene symbols ID having p-value < 0.05, therefore further filtering is needed. Completed this procedure, a list of the gene symbol is extracted from the dataframe and uploaded on the online software of EnrichR, available at the *maayanlab.cloud* website. After the insertion of the list of gene symbol IDs, the main categories taken in consideration for the enrichment were: *Transcription, Pathways, Ontologies and Diseases/Drugs.*

* 1. ***STRING DB***

As the final step of the project, the list of gene symbols previously used in the enrichment analysis was, first of all, sorted according to the p-value in an ascending way, and later the top 150 rows were selected and translated to their respective UNIPROT ID with the aim of subjecting them to STRINGdb analysis, in order to find possible relationships and insights that can be used for future studies. The software itself is able to produce highly informative networks using the provided list of IDs and the information coming from many different databases, with the possibility of choosing the organism of preference.

In the end a network of 133 elements was built, showing how each node (protein) interacts with its neighbours, their relationship and if there are any solitary nodes.

1. ***RESULTS***
   1. ***PCA***

Following the procedures applied in the Principal Component Analysis step (*see Methods*) plots related to the sample clustering and distribution were produced. Regarding this last one, everything seems to be fine, no errors nor unusual elements or modification were found on the other end the PCA plot, when produced, presented one only issue.

Immagine che contiene testo, schermata, diagramma, linea

Descrizione generata automaticamente

Figure 13, Plot deriving from the PCA phase, here almost two perfectly divided groups can be seen except for two datapoints in the bottom part of the graph

In *Figure 13* we are able to see how the samples used can be separated in an almost perfect manner taking into account the tissue type to which they belong, the only exception is the presence of two datapoints, later confirmed to be GSM1258081 and GSM1258082, which seem to have been mixed up during the process, this hypothesis was proposed due to them being the only anomalies in the procedures. A better visualization of the “overlapping” between the two groups can be seen through *Figure 14:*

Immagine che contiene testo, diagramma, schermata

Descrizione generata automaticamente

*Figure 14, PCA plot showing a better clustering divisions of the two groups, in it the overlapping between the two of them can be easily seen*

* 1. ***K-MEANS AND HIERARCHICAL CLUSTERING***

After completing the PCA procedure, different clustering algorithms were applied on the sample study, leading to the correct grouping of each element to its respective class. The results can be seen in *Figure 15* and *Figure 16*, where the labelled samples do not present overlapping.

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

Figure 15, K-means plot obtained in step 3.2., it presents the two classes and the labelled samples belonging to each one, together with the GSM ID of each sample

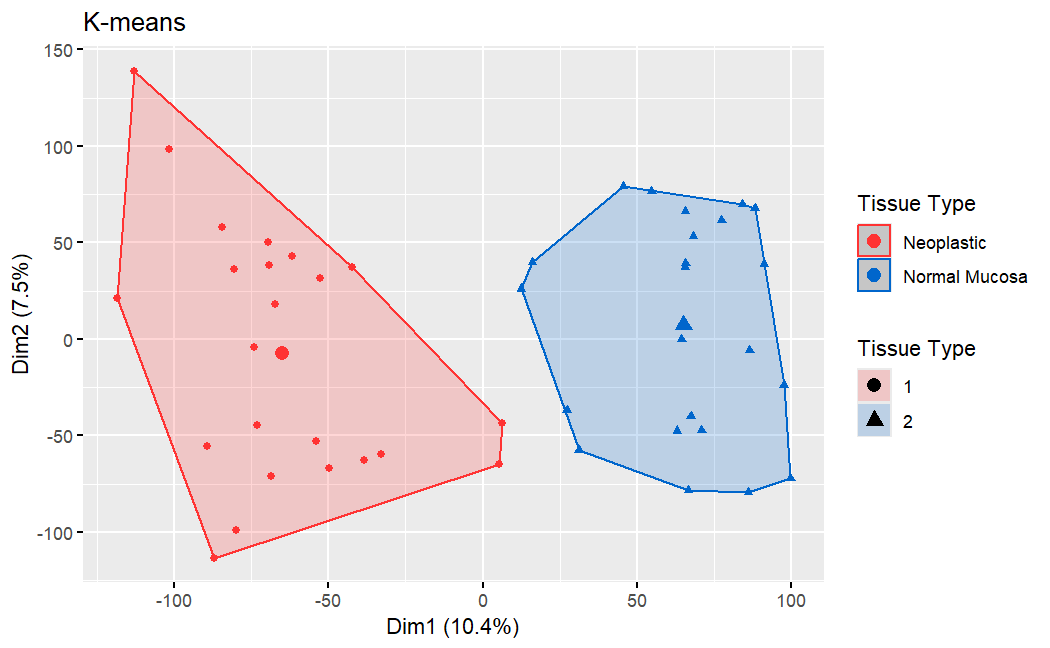
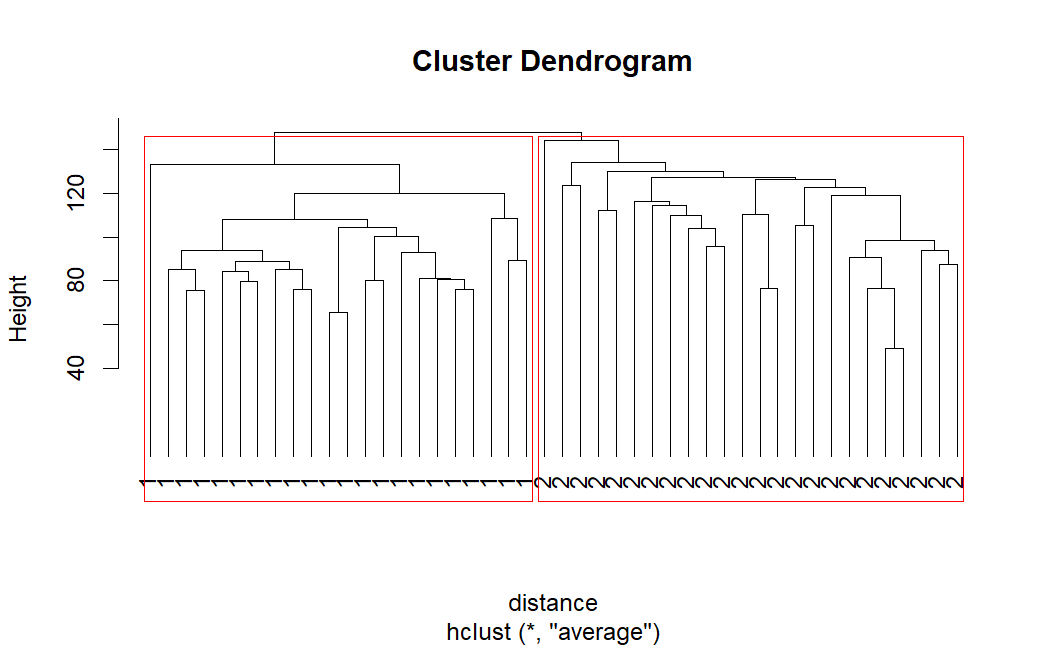


Figure 16, K-means algorithm plot showing the frame of the clustering groups, no overlapping is in sight this time

Using the algorithms applied in the previous step we were able to effectively produce a perfect division of the two groups without the presence of any errors or anomalies, solving the initial problem presented in the PCA algorithm. In *Figure 15,16* the red cluster is characterized by the Neoplastic Tissue, meanwhile the blue is representative of the Normal Mucosa tissue type.

Immagine che contiene testo, diagramma, linea, Rettangolo

Descrizione generata automaticamenteSubsequently to the application of the K-means algorithm, the same test/trial was carried out using the Hierarchical clustering methodology, which resulted also in this case in the perfect subdivision of the two sample groups (*check Figure 17 and Figure 18*)

Figure 18, A better visualization of the previous dendrogram with the branches coloured according to the group: B for Neoplastic and R for Normal Mucosa, and the leaves having their GSM ID

Figure 17, Dendrogram plot produced with the hierarchical clustering algorithm, the two marking in red are used for visualization purposes in order to distinguish between the two groups

* 1. ***HEATMAP***

After obtaining the modified distance matrix using the already available data from the clustering phase, an heatmap was produced:  
Immagine che contiene testo, modello, linea, punto

Descrizione generata automaticamente

Figure 19, Heatmap presenting the comparison between each sample

From the plot we can see that with an inner-group comparison, the values obtained are quite low, on the other end doing an inter-group comparison leads to high value of correlation between the samples compared, maybe underlying possible relationships between different sample’s interactions that may be targeted for further and more specific analyses, while possibly focusing at the gene-expression level to denote key differences or common points/grounds that can be exploited for future studies.

* 1. ***RF***

The application of this methodology resulted in a very clustering that was able to increase its own accuracy also after a set number of trials, as we will be able to see later. In addition to that, from the graph presented in the *Methods* section, we are able to notice which are the main variables that are able to contribute to the reduction of “impurities” (in this case the uncertainty) of the model for the prediction, the top three among these are: “ILMN\_2202948”, “ILMN\_1699357”, “ILMN\_1803312”.

We also obtained the Index Graph, representing how the error rate changes when more trees are built, we can see how with the increase of the tree there is also a correlation with the reduction of the error.

* 1. ***LDA***

During part 3.6. procedures related to the Linear Discriminant Analysis were applied, one of the first plot produced was that related to the quality evaluation of the LDA algorithm in both the training and test group.

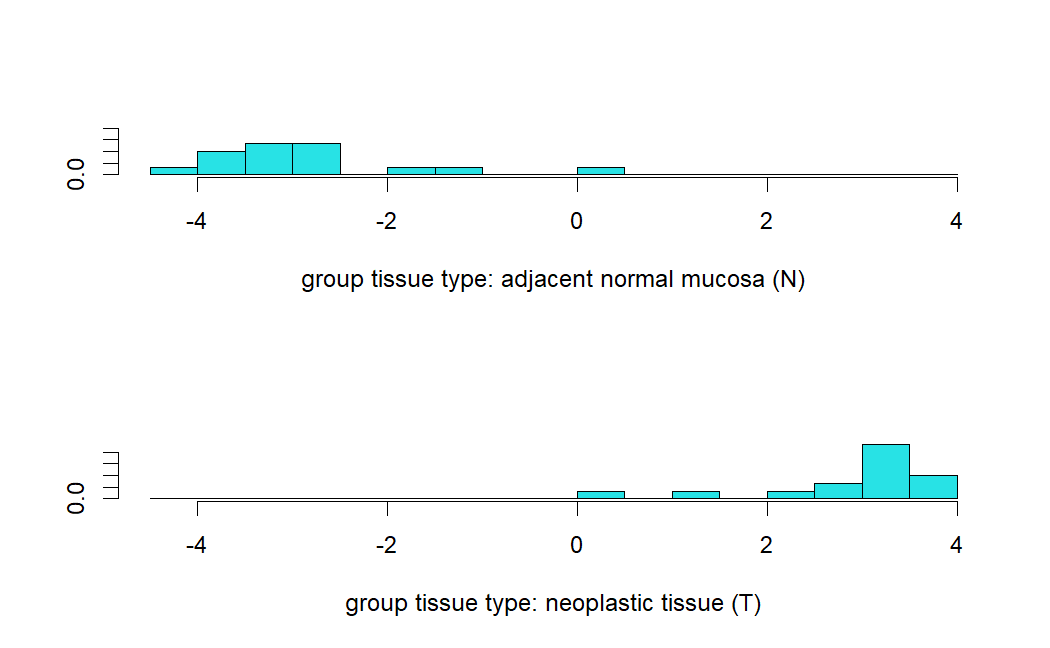


Figure 20, LDA quality evaluation in both [UP] Normal Mucosa group and [DOWN] Neoplastic group, the X axis represents the division point between the two groups.

A fairly good classification was done by the LDA with the presence of few outliers mainly present in the control group. A possible explanation, also in order to reconnect to what was already noted from previous results, is that the outliers represent the nodes that were later fixed using the k-means clustering algorithm in step 3.3., therefore possibly hinting again to a certain relationship between these samples, considering also the possibility of erroneous data collection and elaboration. Further analysis and consideration shall be taken if the aim is that of solve this doubt.

Additional evaluating operations regarding the clustering/grouping of the sample while using LDA were executed:

Immagine che contiene schermata

Descrizione generata automaticamente

Figure 21, LDA clustering plot, in [blue] we have the samples related to the control group (Normal Mucosa), in [green] we can see the samples belonging to the Neoplastic tissue group

As we can see also in this case an almost perfect grouping of the samples is done except for the one that are already known to cause some anomalies, referring, for example, to data point n.32 in the bottom right part of the graph, and using the information we already know we have the possibility to identify the outlier as the sample having the ID GSM1258090.

Additionally, in order to understand how the LDA faired in this test, the trade-off between specificity and sensitivity was calculated, with a final AUC value of 0.88, demonstrating that the model itself does a good job when it comes to classify examples in this particular situation, not perfect but good enough*.*

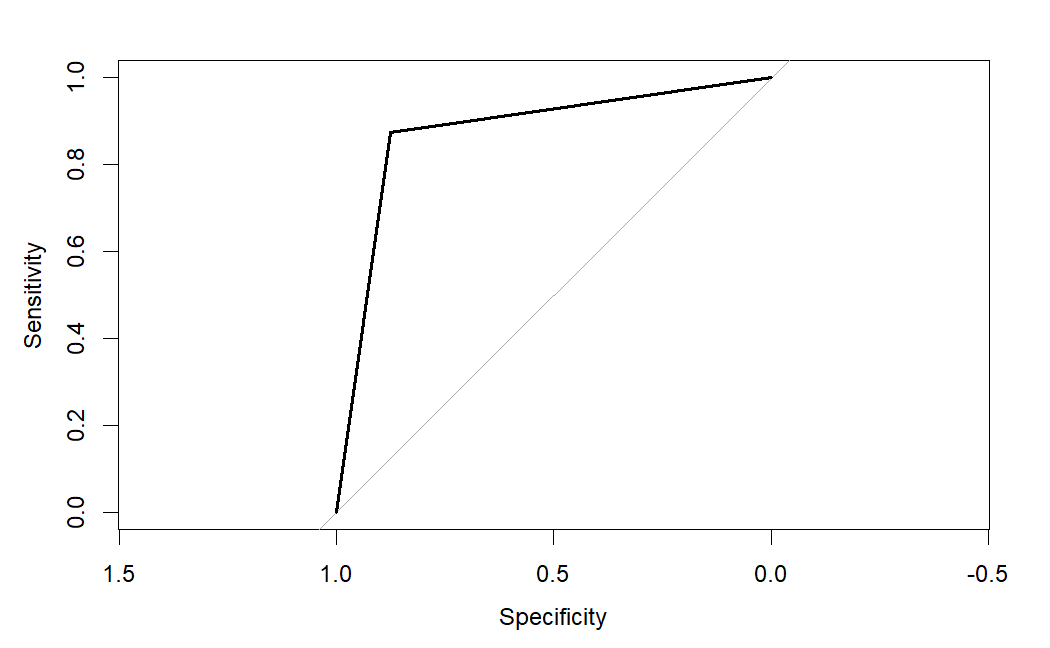


Figure 22, ROC curve and representations of the AUC of the LDA methodology

Following the AUC calculation, the next phase of evaluation concerns the comparison of the models used (RF and LDA) in terms of Accuracy. Using model prediction functions, the following results related to their efficacy were obtained (*check Figure 23*):

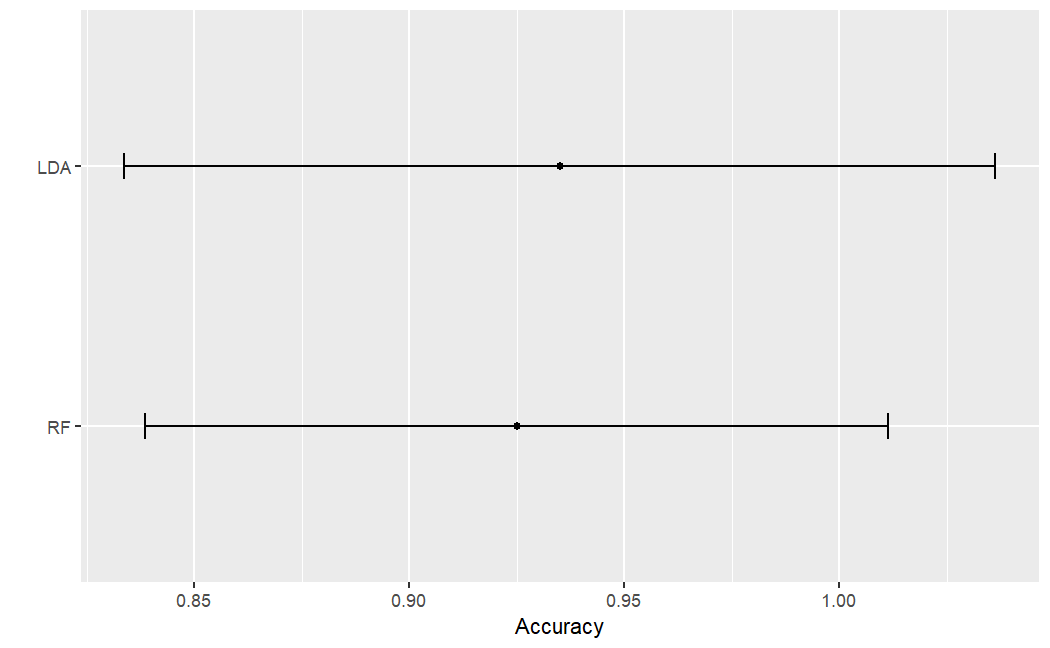


Figure 23, comparison between the application of LDA against RF, as we can see the accuracy of LDA is much higher than expected, with both its mean and IQR being grater than the RF model

By looking at the graph we are able to see that the accuracy of the LDA model is still high, but it does not seem to be due to the presence of overfitting of the training dataset, for the RF, instead, the accuracy seems to be much more in standard ranges, the causes of this event are still no known in our case.

As a supplementary operation, the application of the CARET methodology was applied so as to se how both the models would act in presence of more trials. The results showed that in an opposite respect to what we have seen before, with the implementation of more trials the best model to apply is Ranndom Forest instead of LDA (*see Figure 24*):

Immagine che contiene schermata, diagramma, linea, testo

Descrizione generata automaticamente

Figure 24, Plot showing the comparison of the Accuracy of LDA vs RF, in the case of CARET application

* 1. ***LASSO/RIDGE REGRESSION***

During the 3.7 phase of the project, different results and plots were obtained when the LASSO methodology was applied, starting off with the plot needed in order for us to understand which was the best λ value to use for producing the lowest MSE value, with λ being equal to -2.46

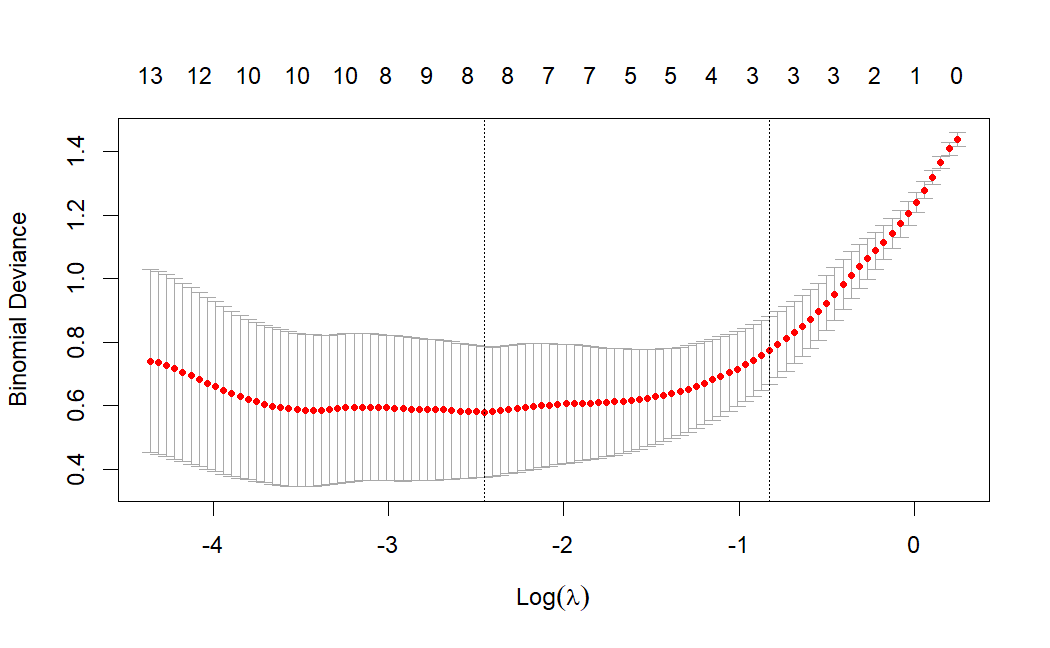


Figure 25, plot showing the lambda value such that it is able to minimize the binomial deviation, obtained in the LASSO analysis

Afterwards a plot representing the relationship between regularization parameter specific for LASSO and the model accuracy was drawn. From that graph we can gain some insights regarding how the reg. parameter can be inserted in the model while also preserving the quality of the classification, showing also that with the increase of its value the model itself probably becomes too much simplistic and is not able to work as well as before (*look up Figure 26*).

Immagine che contiene testo, Diagramma, linea, diagramma

Descrizione generata automaticamente

Figure 26, Plot representing the relationship between the accuracy of the model and the value of the regularization parameter to insert in it, increasing the parameter means lowering the accuracy of the model

As the final step in this stage of the experiment, the model obtained through the application of the LASSO methodology was compared to those of RF and LDA, leading to the LASSO model resulting as far worse to apply to the initial dataframe respect to its two colleagues.

Immagine che contiene schermata, testo, linea, diagramma

Descrizione generata automaticamente

Figure 27, Comparison plot between the three methodologies used for clustering/classification procedures, when implementing the LASSO model, the accuracy of the classification greatly decreases in comparison to RF or LDA

* 1. ***RSCUDO***

Concluding the LASSO classification procedure, the RSCUDO methodology was applied. The first results obtained showed the clustering of the tissues groups in two main clusters with the presence of sparse datapoints, with some of them still connected, even if on their own. In *Figure 28* we are able to see what was mentioned above, in addition to that, it is clear the existence of some nodes that seem to “trespass” between the domains of the two key clusters, with a higher probability of them being outliers that were considered related to a group to which they are not actually associated with. The following classification was focused on looking up for the Signatures of the tissue’s groups:

***Immagine che contiene schermata, Elementi grafici, design

Descrizione generata automaticamente***

Figure 28, Clustering plot obtained through the application of the RSCUDO methodology, in this representation the two core clusters can be seen together with a few of other sparse elements which seem to intrude between each other’s domains.

A secondary classification procedure was carried out, with respect to the first one, with the objective of further decomposing the dataset, in order to create much more subclusters. The aim here is to see how the algorithm considers the “solitary” points already seen in Figure 28, on top of the presence of possible changes in the network due to these new applied requirements:

Immagine che contiene Arte bambini, disegno, illustrazione, cartone animato

Descrizione generata automaticamente

Figure 29, Clustering plot showing how the RSCUDO algorithm decompose and recognise the different solitary nodes already seen in the previous graph, here we are able to see the subgroups perceived by it

* 1. ***ENRICHMENT***

After obtaining the gene list the result underwent enrichment through the use of the online platform of EnrichR, as specified in the *Methods section 3.10.,* out of the ~13300 genes, only a fraction of them were found to have a description in the online software, with a number around the ~9000 genes. In the end the enriched genes were around ~4000.The results coming from this step of the analysis presented some associations with:

* Translation
* Influenza infection
* RNA and Ribosome modifications, formation and binding
* Diamond-Blackfan anaemia
* Belong to the top 500 genes that are downregulated in COVID-19
* Glycosylation-related disorders
* Bladder, Ovarian, Testicular carcinoma
* Other different types of aneamia
* Presence of bulk tissue in the kidney

Quite not what we were expecting, but as stated before these results may be due to the high amount of filtering that was imposed on the genes and the presence of duplicates among them. eventually, with further research, some of the pathway in which with elements were enriched were related to the CRC context, for example:

* Colorectal carcinoma
* Colorectal neoplasm
* Acute myeloblastic leukemia
* Carcinoembryonic antigen
* Muir-Torre syndrome
* High association to the tissue of intestine and uterus
* Gut microbiota beta diversity
* Colon cancer association and intestine epithelial cells
* CL34, HT115, SNU16 and SNUC1 cell lineages of the Large Intestine
* Presence of bulk tissue in the colon
* Association with HCT116 cell line

Of particular interest was the finding related to the Muir-Torre syndrome, an autosomal dominant phenotypic variant of hereditary non-polyposis colorectal cancer, characterized by the presence of sebaceous tumours of the epithelial cells of the intestine, usually accompanied by also colonic carcinoma. The genetic modifications related to the syndrome upbringing are known but still needs to be further refined and could be an interesting point of start for a study.

Immagine che contiene schermata, testo, schermo, software

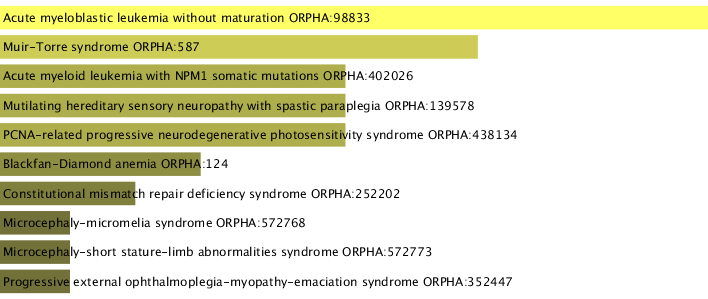
Descrizione generata automaticamente

Figure 30, Orphanet Augmented 2021 section of the enrichment approach applied on the filtered genes, here are the argument presented by the already cited section in which the protagonist is the Muir-Torre syndrome, a form of colorectal cancer, non polypolis, characterized by the presence of sebaceous tumours

In the end both expected and unexpected data were recovered, with some of them with a higher possibility of being used for the treatment of the disease, even though more information are needed for a better characterization of the condition and for better assessment of the weak point that can be used against CRC, even though low relevance to the disease was found from the obtained genes.

* 1. ***PATHFINDR***

With the application of the PATHFINDR algorithm, supplementary information regarding insights on the enrichment of the genes used before, it showed much similar results to those obtained using the EnrichR instrument. Such an example is present in Figure 31, where the plot introduced showed high level of relationship between the gene list inserted and their respective relevance in ribosomal functions, modification and structural features, which were also present in the previous phase.

Immagine che contiene testo, schermata, numero, Carattere

Descrizione generata automaticamente

Figure 31, Enrichment plot that presents the results coming from the base KEGG database used in PATHFINDR, here the results are highly correlated to the one found in the previous phase using the EnrichR online instrument

The results obtained from this phase mirror those coming from the EnrichR online tool, with the only exception that with the produced plots we are not able to see any relevance to the Colorectal region from which the tissues were obtained. Additional plots were produced:

Immagine che contiene testo, schermata, numero, Carattere

Descrizione generata automaticamente

Figure 32, Enrichment ploit produced by PATHFINDR in which the genes are enriched with respect to Biological Process information

Immagine che contiene testo, schermata, numero, Carattere

Descrizione generata automaticamente

Figure 33, Enrichment plot related to the Reactome database, obtained through the PATHFINDR package

* 1. ***STRING DB***

Once the gene list from the merged dataframe was obtained, the first result obtained from this step was a list of the respectively translated genes into a suitable format for STRING DB, that being the respective protein name for each respective gene. The final completed list was then submitted into the online software, with also the specification of the organism of interest (Homo Sapiens). In the end a network was generated showing how the elements are connected:

Immagine che contiene fiore, schermata, luce

Descrizione generata automaticamente

Figure 8, STRINGDB- derived plot presenting the main connected cluster with few sparse solitary nodes

The list of the main connected components present in the central cluster was obtained manually and was later put also under enrichment for further investigation, those being: CEP55, UBE2T, MND1, CKS2, ASPM, CCNA2, EXO1, CCNB1, NCAPG, TPX2, MCM6, UBE2C, BUB1, GINS2, RFC3, MCM10, KIF2C, KIF4A, TRIP13, RACGAP1, CENPW, SKA3, FAM83D, CDK4, GMPS.

Results showed enrichment in gut microbiota association, together with melanoma, gallbladder and renal cortical glands diseases, connection to Huntington disease, relationship with neurodegenerative disease, liver carcinoma and Goldberg-Shprintzen megacolon syndrome.

***5.0. FINAL THOUGTHS***

Since the beginning this study aimed to focus on the comparison of different methodologies to classify the most correctly as possible samples coming from tissues known to be related to the colorectal cancer affliction, while considering their effectiveness and efficiency, trying in the meantime to avoid errors and biases introduction as much as possible, considering also each methodology behaviour with the “perceived” presence of outliers. We now know that their quality depends on many factors, such as the number of trials introduced the number of samples, in example when faced with multiple tests, the Random Forest methodology showed the best accuracy mean and range, in the case of the single run, instead, LDA reign over the others, in the meanwhile LASSO was the single methodology which managed to score the lowest accuracy. In addition, RSCUDO clustering showed how the samples behaved and their relationship between each other, clearing the path for the further analysis that took place throughout the final sections of the project

Even though the data in the end was heavily filtered, the information obtained from the enrichment approach were still significant and further attempts may be ideal for better assessment of possible study-target for CRC treatment. Unfortunately, some information are still missing, such has been seen when using the STRING database with the presence of some stand-alone nodes, demonstrating the need for more advanced characterization is required.