**BTP Report**

Name: Kaki Hephzi Sunanda

Roll Number: 210150018

**Objective**

The objective is to classify colorectal carcinogenesis samples based on gene expression profiles obtained from microarray integration (GPL570) and normalized COAD TCGA gene expression data. The primary objective is to identify the minimal number of gene features required to classify samples as cancerous or non-cancerous at different certainty thresholds (e.g., 90%, 95%). The obtained genes are later used classify the COAD TCGA gene expression data.

**Dataset**:

For the task three datasets were given:

1. The normalized gene expression profile of colorectal carcinogenesis that was determined using GPL570 microarray integration.
2. The corresponding phenotypic information of the samples in the preceding dataset.
3. The normalized COAD TCGA dataset.

**Data Exploration:**

The normalized gene expression profiles data consists of 7830 genes and 1339 samples. The samples are classified into one of the four phenotypic classes (Adenoma, Cancer, IBD, Normal).

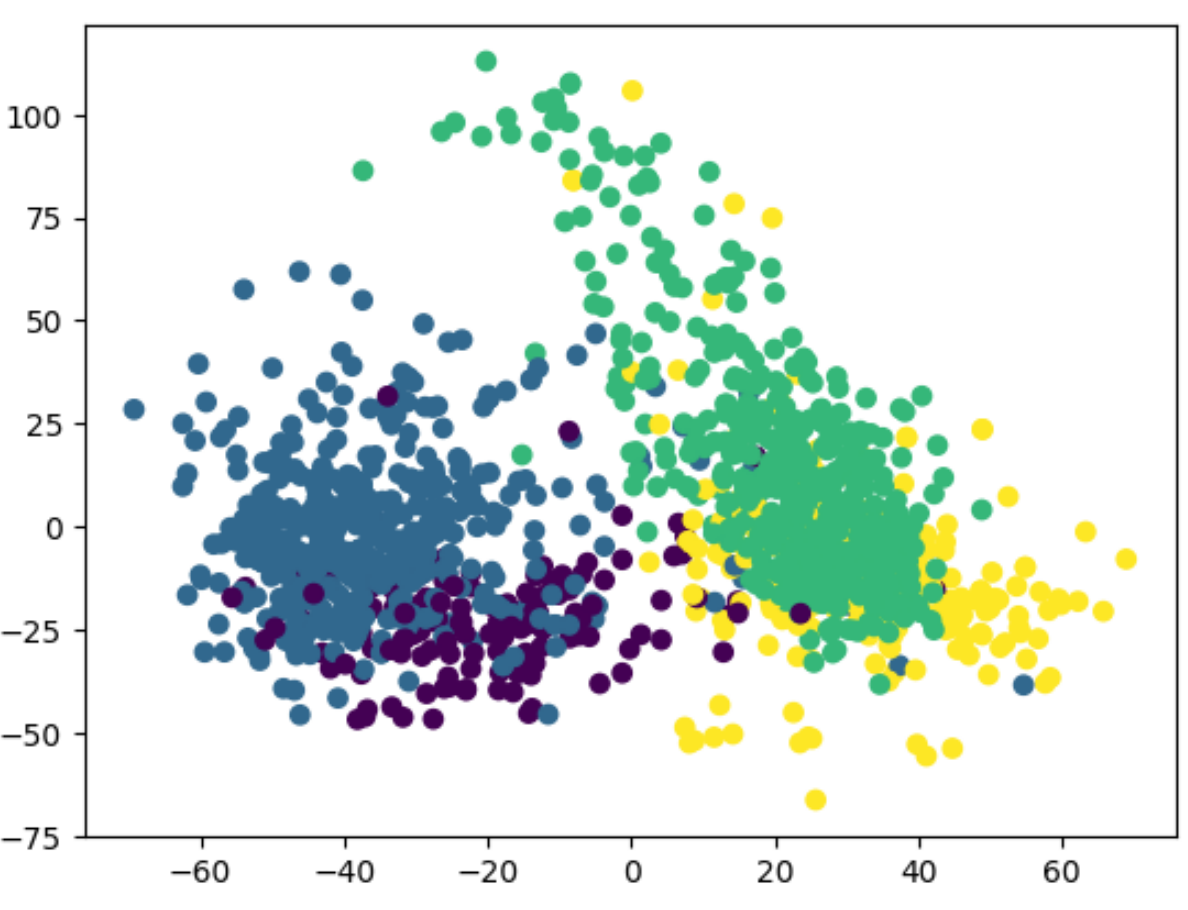
A histogram and a density plot have been used to visualize the distribution of the gene expression profiles across all samples and genes.



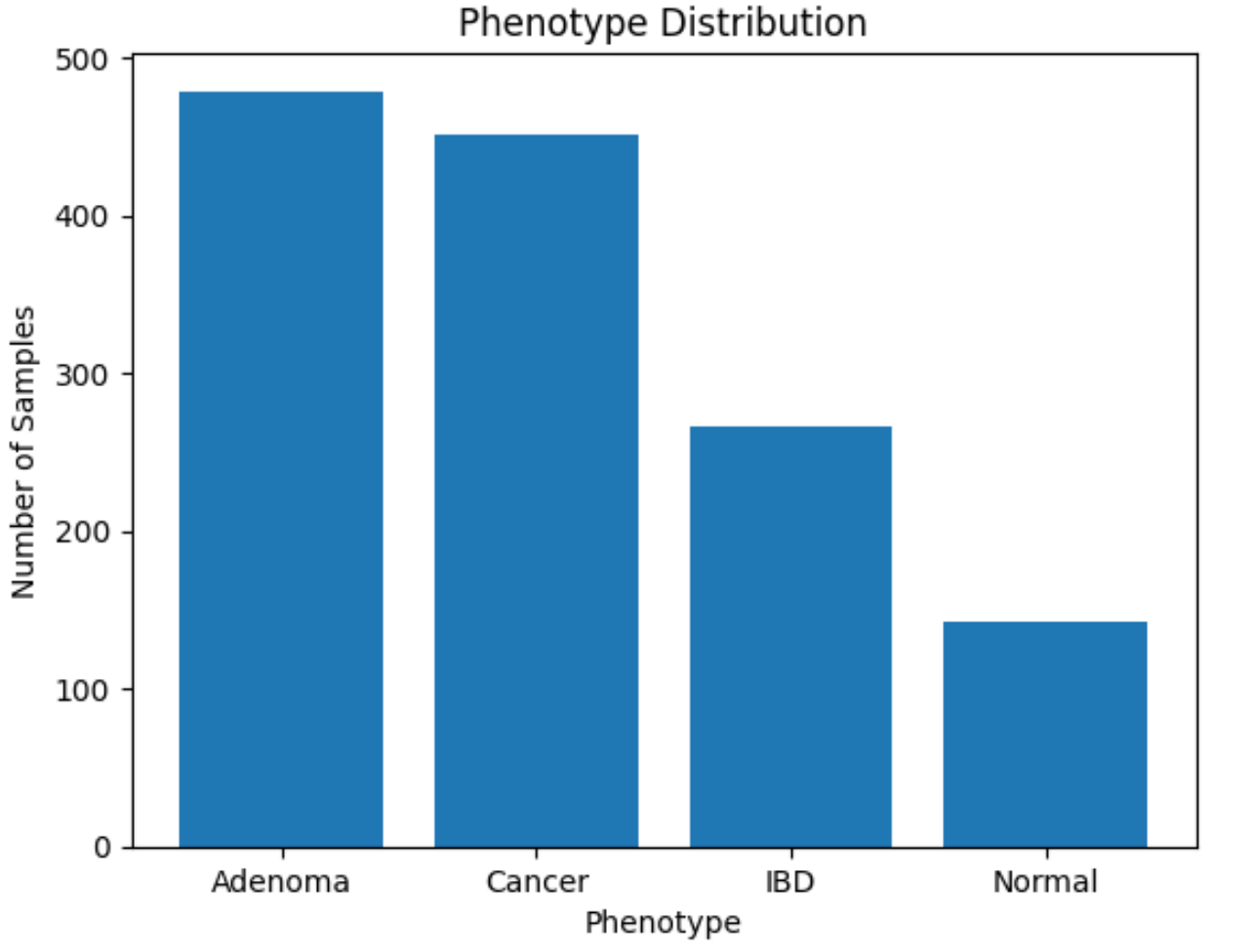


From the plots, it can be observed that the distribution approximates to a gaussian distribution with a slight positive skew. The skew could indicate the relatively high expression levels of a few genes.

A PCA plot was used to visualize the sample clusters in two dimensions.



The samples do cluster into several clusters, that may suggest gene expression profiles independence across the different phenotype classes. There is a significant overlap between some clusters indicating that feature selection may be required to completely differentiate the classes.



The above bar plot shows the distribution of samples across the different phenotypic classes. Adenoma, Cancer, IBD and Normal phenotype classes contain 479, 451, 266 and 143 samples respectively. The class imbalance present may influence the classification performance.

**Data Preprocessing:**

The following steps were performed for data preprocessing**:**

* *Transposing the Gene Expression Data*:  
  In the GPL570 normalized gene expression dataset, the rows corresponded to genes and the columns represented samples. Since the feature selection and classification models require samples as rows and genes as columns (representing the features for the feature selection and classification model), the dataset was transposed such that rows now represent individual samples and columns represent the gene expression profiles of the samples. The sample names were set as the index for proper identification during merging.
* *Handling the Phenotypic Information*:  
  The second dataset, consisting of the phenotypic information of the samples, consisted of rows representing samples and columns containing various metadata, including a column with phenotype information (Adenoma, Cancer, IBD, Normal). For this dataset, the sample names were again set as the index to ensure consistency when merging with the transposed gene expression dataset.
* *Merging Datasets*:  
  To create the final dataset for feature selection and classification, the phenotype information was merged into the transposed gene expression dataset. This merge was performed based on the sample names (set as the index), ensuring that each sample’s gene expression profile was linked to its corresponding phenotype. The final dataset used for feature selection and classification contained:
  + Rows as Individual samples.
  + Columns as Gene expression values for each gene across all samples and the corresponding phenotype (cancerous or non-cancerous).

**Methodology:**

One of the objectives was Feature Selection. Reducing the number of features can simplify the model and enhance the performance of the classification model by focusing on the most informative features.  
Since the dataset is labelled, supervised feature selection methods have been used. These methods have assessed the relevance of every gene concerning the target phenotypic class labels. The following types of Feature selection methods have been used:

1. *Filter-based methods*:

The filter-based feature selection methods evaluate each feature importance using a statistical measure to rank their relevance to the target label / class. Common filter methods include: Information Gain, Chi-Square Test, Fisher’s Score, ANOVA F-Test.

1. *Wrapper-based methods*:

The wrapper-based feature selection methods evaluate subsets of features based on model performance. The model iteratively selects features and evaluates them through a classification algorithm. Some examples include: Forward Selection, Backward Selection, Recursive Feature Elimination (RFE), Exhaustive Feature Selection.

1. *Embedded methods*:

The embedded methods integrate feature selection as part of the model training process. The model itself identifies the most important features. Some examples are: Random Forest Feature Importance, Regularization (Lasso).

The given gene expression data contained feature redundancy. To eliminate any negative effects on the classification performance, a Spearman correlation matrix was computed to evaluate the pairwise correlation between genes. Multiple thresholds were tested to determine the optimal level of correlation removal.

The threshold of 0.80 was selected. At this threshold, informative genes were retained and redundancy was removed. This processing step dropped 1055 genes.

Stability checks were also incorporated to evaluate the robustness of feature selection and classification methods. Each feature selection method was run multiple times to identify features that consistently appeared across the iterations. This ensured that the selected features were robust.

Similarly, classification models were also evaluated over multiple iterations. This iterative approach enhanced the reliability of the results.

Three classification models were tested to evaluate the performance of the selected features: k-Nearest Neighbors (kNN), Support Vector Machine (SVM), and Random Forest.

**Results:**

After applying the ANOVA F-Test for feature selection, there were a total of 27 genes that had over 50% frequency in the iterations.

The classification performance was evaluated using Random Forest on the selected genes. The metrics obtained were as follows:

|  |  |  |
| --- | --- | --- |
| **Metric** | **Class: Normal** | **Class 1 : Non-Normal** |
| Precision | 1.0 | 0.96 |
| Recall | 0.87 | 1.0 |
| F1-Score | 0.93 | 0.98 |

After applying the Lasso Regularization for feature selection, there were 16 informative genes and the classification metrics obtained were as follows:

|  |  |  |
| --- | --- | --- |
| **Metric** | **Class: Normal** | **Class 1 : Non-Normal** |
| Precision | 0.96 | 0.99 |
| Recall | 0.96 | 1.0 |
| F1-Score | 0.96 | 0.99 |

After applying the Random Forest Selection, the classification metrics are:

|  |  |  |
| --- | --- | --- |
| **Metric** | **Class: Normal** | **Class 1 : Non-Normal** |
| Precision | 1.0 | 0.98 |
| Recall | 0.82 | 1.0 |
| F1-Score | 0.90 | 0.99 |

**Conclusion:**

This project classifies colorectal carcinogenesis samples using gene expression profiles by identifying a minimal set of genes through feature selection techniques. ANOVA F-Test identified 27 genes with high stability, Lasso Regularization reduced the gene set to 16 genes, and Random Forest Importance selected features that could enhance the classification performance. Among the tested classification models, k-Nearest Neighbors (kNN), Support Vector Machine (SVM), and Random Forest, the Random Forest and Lasso model demonstrated best results for the “Non-normal” class, achieving an F1-score of 0.99. Future work could focus on extending the methodology to classify all phenotypic classes, and integrating additional datasets, such as transcriptomics or proteomics, to improve classification accuracy further.