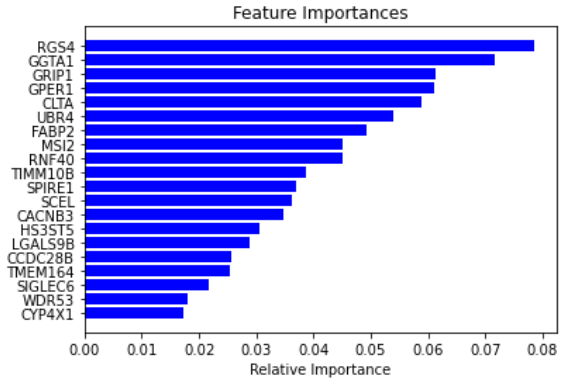
Feature Analysis Project

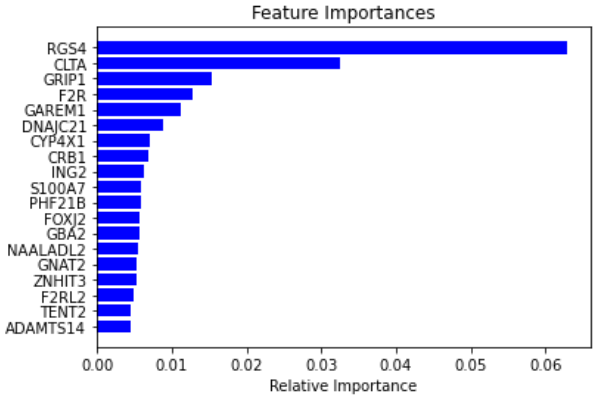
By: Atoosa Ayazbkahsh

Feature evaluation

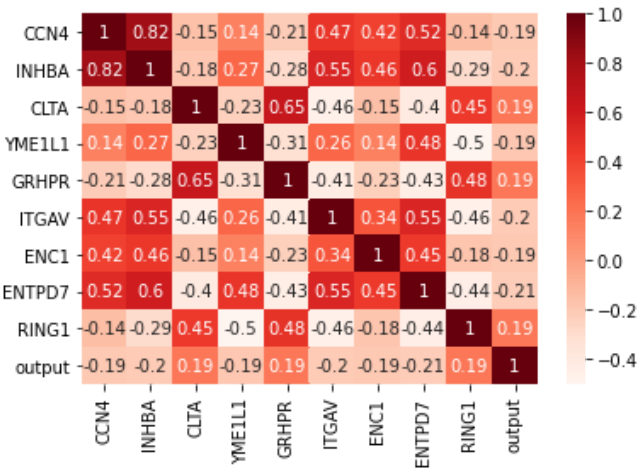
The given data was 588 samples with 19566 features which were gene names, and 613 labeled data samples were in the TCGA-BRCA.pheno file. However, only some of the samples in TCGA-BRCA.pheno were included in TCGA-BRCA.htseq_fpkm-uq_gene_name, I had to subset and realign data.
With *columns.intersection* I extracted mutual columns, which were features after I transposed the feature data frame. Furthermore, I relabeled data with integers which 1 represented “Tumor” and 0 denoted “Normal”. Finally, I ended up with 588 labeled samples.
I trained my data with **Decision Tree Classifier** to discover relevant features by its *feature_importances_* attribute. The top 20 features are demonstrated in the bar chart below.



I also tried this method with **Random Forest Regressor** using the *feature_importances_* attribute, and the top 20 features are demonstrated in the bar chart below.



Moreover, I implemented **Pearson Correlation** to discover how correlated features were with the results. The absolute value of the correlation ratio was calculated, and the output list was sorted. The top 10 features were selected to demonstrate in the following heatmap figure.



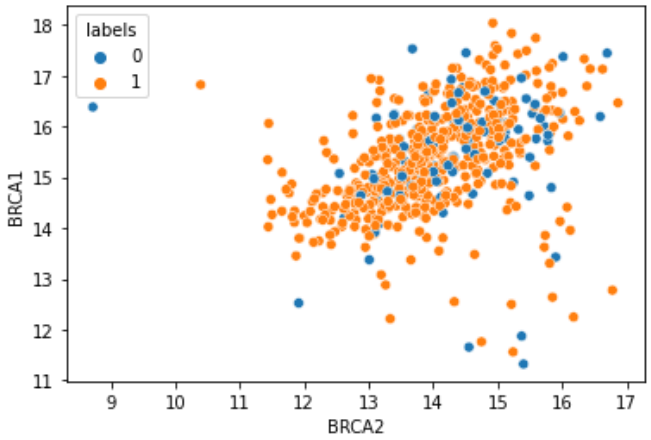
Also, 100 top features were saved in the **feature_analysis_AtoosaAyazbakhsh.csv** file. The absolute value of the correlation ratio was added to the data frame, but extracting the actual correlation ratio was also possible, and its code was written.

Data Clustering with Selected Features

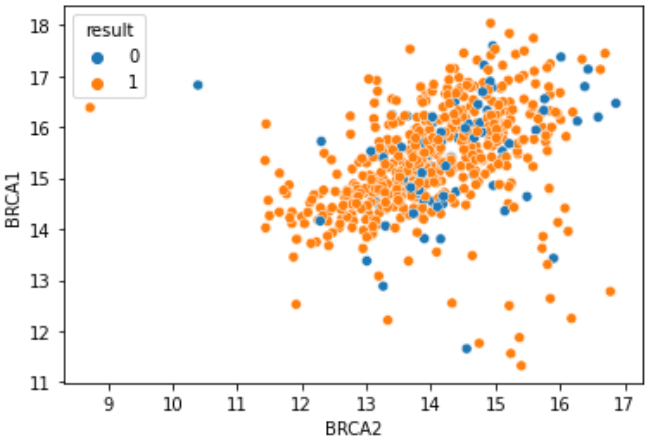
50 top features were selected to perform clustering on the data. **Spectral Clustering** and **KMeans** algorithms were implemented. Spectral Clustering could bring more accurate results due to its usability for data with few clusters. (“2.3. Clustering — scikit-learn 1.2.2 documentation”, n.d.)

Conclusion

CLTA feature was one of the mutual features among others, and it was referenced in some papers. (María G. C. Navarrete-Bernal, n.d.) (Gong, n.d.) It is questionable that BRCA1 and BRCA2, which were famous factors in breast cancer (Lynch JA, n.d.) had little correlation with output. Even visualizing these two features with corresponding labels could not help in finding a relationship between breast cancer and these two features. Visualized data is demonstrated below. The right figure is the actual labeled data, and the left one is the spectral clustering algorithm predicted result.



Spectral Clustering



Actual Data

References

Gong, G. (n.d.). Circular RNA circ_0084927 regulates proliferation, apoptosis, and invasion of breast cancer cells via miR-142-3p/ERC1 pathway. *National Library of Medicin*. 4120-4136

Lynch JA. (n.d.). Genetic tests to identify risk for breast cancer. 10.1016/j.soncn.2015.02.007

María G. C. Navarrete-Bernal. (n.d.). Biological Landscape of Triple Negative Breast Cancers Expressing CTLA-4. 10.3389/fonc.2020.01206

2.3. Clustering — scikit-learn 1.2.2 documentation. (n.d.). Scikit-learn. Retrieved June 12, 2023, from https://scikit-learn.org/stable/modules/clustering.html#spectral-clustering