# Machine Learning with p53 Gene Somatic Mutations

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

**Rupal Saini**

Spring, 2021.

Technical Project Team Members

N/A

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Dr. Nada Basit, Department of Computer Science

# Capstone Research: Clustering and Random Forest

## Abstract

Clustering is a type of unsupervised learning that does not utilize the ground truth. It is a way to group unlabeled samples. Similarity is measured in the data and then clusters are formed where the data points in each collection have some like features that caused the algorithm to place them together. Using clustering, the ground truth of a dataset can be derived. This ground truth can in turn be used to make predictions based on the similar features. The following research walks through this process with a database of p53 gene somatic mutations in human tumors and cell lines. The motivation behind this research is to learn more about machine learning algorithms and create a foundational basis as the world progresses towards a primarily ML environment.

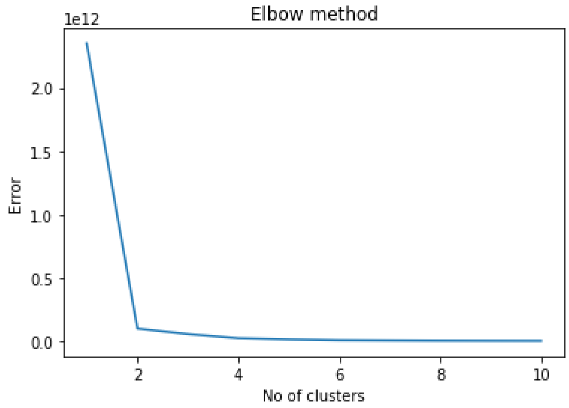
## Methods

The first step of the process was to determine which dataset to use for this research project. I was interested in bioinformatics, so it was ideal for me to use a dataset with some biological relationship. This ended up being a dataset on p53 cancer cells. Initially, I had decided to use the p53 Mutants Data Set in the UCI Machine Learning Repository (Lathrop, 2010), however this dataset did not have descriptive attribute information which made it difficult to understand. I then moved to a Database of p53 gene somatic mutations in human tumors and cell lines (Hainut et al., 1997). This dataset, on the other hand, did not have a ground truth attribute which would describe whether or not the tumor was cancerous. This is where the idea of doing clustering instead of directly going to a machine learning algorithm came about.

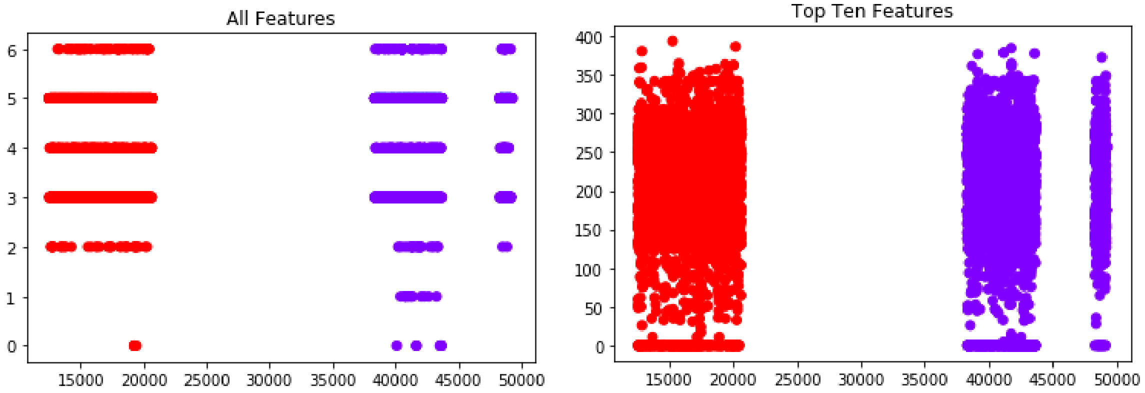
The most difficult part of this research project was formatting the data in a way that could be easily read by the feature selection and machine learning algorithm, also known as preprocessing. To format the data, a label encoder was utilized (“sklearn.preproccessing”, n.d.). This normalized the labels and transformed non-numerical labels to numerical labels so that the algorithm would be able to process the data. This was done to every attribute with non-numerical data.

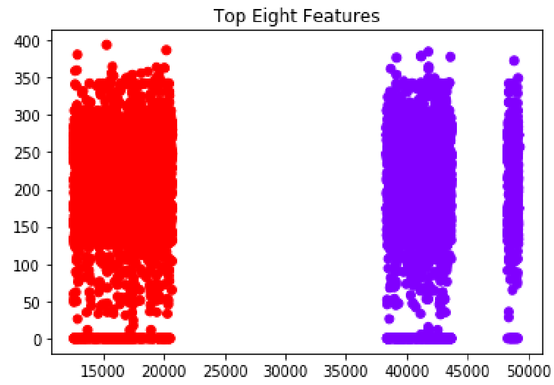
Next, feature selection (Brownlee, 2014) was performed. The method chosen was the Chi-Squared statistical test. This test belongs to a class of filter methods, a process to assign a score to each feature and then rank the features by that particular score (Paul, 2020). The Chi-Squared test was run on all of the attributes and the scores were printed out. The higher the score, the more important the attribute (“sklearn.feature_selection”, n.d.). The following is a list of the features listed out from most important to least important: Mutation_ID (column 0), Morphology (column 16), Sub-topography (column 15), Topography (column 14), Topo_code (column 17), Putative stop (column 12), Mutant_codon (column 5), Codon (column 3), WT_codon (column 4), Description (column 6), WT _AA (column 8), Source (column 13), Mutant_AA (column 9), CpG (column 7), Frameshift (column 11), Splice (column 10), Location (column 2), Type (column 1).

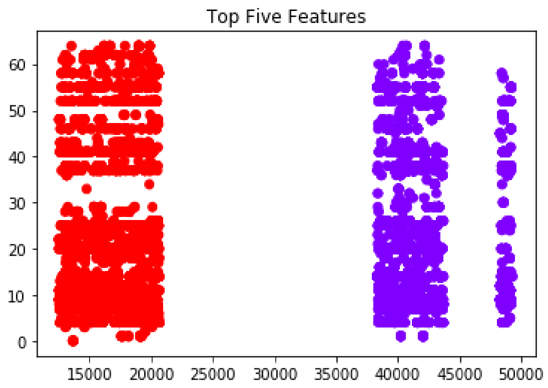
After feature selection, a clustering algorithm was researched. I decided to perform k-means clustering. To begin, the optimal number of clusters for the dataset needed to be determined. I knew that I wanted the dataset to be split into two clusters to represent cancerous and benign, however, I wasn’t sure if that would be efficient. To get around this issue, the elbow method was performed (Dhiraj, 2019). This would plot a graph between the number of clusters and the error value corresponding to that. Wherever the shape of the elbow was formed was determined to be the optimal number of clusters for the dataset. The elbow graph is shown below:



It is evident that the number of clusters the dataset should be split into, is two. This is in accordance with the prediction that I had made. From here, k-means clustering was carried out using a few lines of python code. This was done in a few ways - with all of the features, the top ten features, the top eight features, and the top five features. As stated earlier, the top features were determined through feature selection and the Chi-Square test. The following are the cluster graphs for each k-means algorithm carried out:



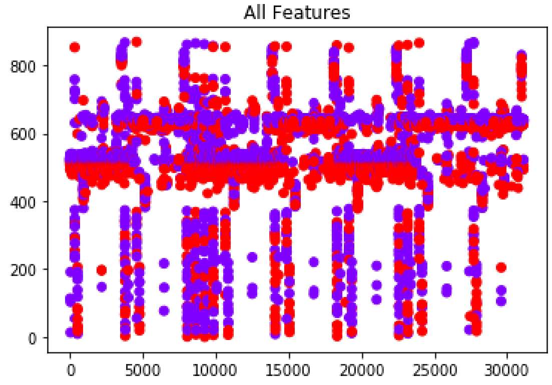




The last portion of this research assignment was to perform the Random Forest machine learning algorithm using the data from the clustering as the ground truth (Navlani, 2018). What I did is add another column to the data representing which cluster each sample belonged to. The ground truth in this case was taken from the k-means clustering results that used all features. The data was split into training and testing sets. 70% of the data was allotted for training and the remaining 30% was separated for testing. A gaussian classifier was made and the model was trained using the ground truth data that was already given. Then, predictions were made for the ground truth for the testing data, using the trained model. After this, the accuracy of the model was calculated. It came out to be a 1.0 accuracy, meaning the trained model perfectly predicted which cluster the testing data belonged to.

An issue that arose was the difficulty of understanding whether or not the Random Forest algorithm was working correctly. It was assumed that it wasn't because usually datasets do not give an accuracy as high as was found for the Database of p53 gene somatic mutations in human tumors and cell lines. To further determine if this was the case, Random Forest was performed a couple more times on other datasets. The first of which being the original p53 Mutants Data Set in the UCI Machine Learning Repository, as this included the ground truth already. Next, Random Forest was performed on the same p53 Mutants Data Set with clustering results as the ground truth. Finally, it was performed on the very well known Iris dataset (Fisher, 1988) to determine if the results of the previous findings were significant.

The process of performing the Random Forest algorithm on the p53 Mutants Data Set in the UCI Machine Learning Repository was very similar to what was done so far. First, feature selection was performed with the Chi Square test, then clustering, and finally the algorithm was used to make predictions. The top ten features of the UCI dataset turned out to be columns 2522, 2972, 3580, 4477, 4616, 4105, 4476, 2548, 4819, and 4098. As stated earlier, this dataset did not have in-depth attribute information, so the columns are being referred to by numbers instead of names. After this, clustering was performed to see where the data points would lie if there was no information on the ground truth given. The graph is below:



It can be seen that the clusters are somewhat overlapping. The purple points belong to one cluster, while the red points belong to another. It is not clear which cluster means that the p53 mutants are active or inactive. The results from the clustering were later used as a ground truth for the Random Forest algorithm. It was predicted that the accuracy of the algorithm would decrease due to the overlapping points. However, after using clustering as the ground truth, when 70% of the samples were trained and the other 30% were tested, the Random Forest algorithm still gave 98% accuracy in its prediction. When the given ground truth in the p53 Mutants Data

Set was used, the accuracy of the prediction under the same constraints (70% trained and 30% were tested) was 99%.

Because these accuracies were so high, I needed to perform Random Forest on a known dataset with a known accuracy to determine whether the algorithm was working correctly. I decided to use the Iris dataset. With the Random Forest algorithm, it is known that this dataset has a 93% accuracy. After running Random Forest on my machine with the dataset using 70% of the samples for training and 30% for testing, the accuracy was indeed 93%.

This led me to believe that the Random Forest algorithm does work. It can also be concluded that the data in the Database of p53 gene somatic mutations in human tumors and cell lines is extremely well separated. This is the same idea for the p53 Mutants Data Set in the UCI Machine Learning Repository. Meaning, the algorithm itself is not faulty, it is simply the nature of the datasets which allowed the algorithm to make predictions at such a high accuracy. Therefore, it is important to note that all previous findings stated in this paper are valid and correct.

## Reflection

Throughout this research, I have learned a large amount of information on machine learning. I came into this project with very limited knowledge on what machine learning is, how it works, and why it is performed. Now I can answer these three questions and more with a very in-depth approach. By performing feature selection, clustering, and the Random Forest algorithm in my own environment, my understanding was furthered in a very significant way. Prior to this project, I had not even realized that many machine learning algorithms were already written and it is simply a few lines of code that need to be written to actually apply the algorithm to a dataset.

Moreover, it is now evident that the hardest part of machine learning is modifying the dataset in a way that it is uniform. This in itself took me a couple of days to carry out. I am confident that I can use machine learning in the future whether it is in a job or a school project. I am very grateful to have had the opportunity to pursue this and learn more about a very interesting field of computer science with the mentorship of Professor Basit.

## Future Work

Other uses that may come of this research is using the same model to perform random forest on the different types of k-means clustering. Meaning, using feature selection and the ground truth that was found from k-means clustering of 10, 8, and 5 features, how does the accuracy of the model differ? I am also interested in using different machine learning algorithms such as logistic regression, association rule mining, and more to determine which is a more accurate predictor. Further, I can see what difference k-fold cross validation makes on the model.

## Sources

Baronio, R., Ho, L., Hall, L., Salmon, K., Hatfield, G. W., Kaiser, P., & Lathrop, R. H. (2009, September 4). *Predicting positive p53 cancer rescue regions using most informative positive (MIP) active learning*. PLOS Computational Biology. Retrieved November 2, 2021, from https://journals.plos.org/ploscompbiol/article?id=10.1371%2Fjournal.pcbi.1000498.

Brownlee, J. (2020, June 30). *Introduction to dimensionality reduction for machine learning*. Machine Learning Mastery. Retrieved November 2, 2021, from https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/.

Brownlee, J. (2014, October 6). *An introduction to feature selection*. Machine Learning Mastery. Retrieved November 2, 2021, from https://machinelearningmastery.com/an-introduction-to-feature-selection/.

Danziger, S. A., Zeng, J., Wang, Y., Brachmann, R. K., & Lathrop, R. H. (2007, July 1). *Choosing where to look next in a mutation sequence space: Active learning of informative p53 cancer rescue mutants*. Bioinformatics (Oxford, England). Retrieved November 2, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2811495/.

Danziger, S. A., Swamidass, S. J., Zeng, J., Dearth, L. R., Lu, Q., Chen, J. H., Cheng, J., Hoang, V. P., Saigo, H., Luo, R., Baldi, P., Brachmann, R. K., & Lathrop, R. H. (2009, September 22). *Functional census of mutation sequence spaces: The example of p53 cancer rescue mutants*. IEEE/ACM transactions on computational biology and bioinformatics. Retrieved

November 2, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2748235/.

Fisher, R. A. (1988, July 1). *Iris Data Set*. UCI Machine Learning Repository: Iris data set. Retrieved November 2, 2021, from https://archive.ics.uci.edu/ml/datasets/iris.

Hainut, P., Soussi, T., Shomer, B., Hollstein, M., Greenblatt, M., Hovig, E., Harris, C., & Montesano, R. (1997, January 1). *Database of p53 gene somatic mutations in human tumors and cell lines: updated compilation and future prospects*. Oxford Academic. Retrieved November 2, 2021, from https://academic.oup.com/nar/article/25/1/151/1085590.

Lathrop, R. H. (2010, February 9). *p53 Mutants Data Set*. UCI Machine Learning Repository: P53 mutants data set. Retrieved November 2, 2021, from https://archive.ics.uci.edu/ml/datasets/p53+Mutants.

Navlani, A. (2018, May 16). *Sklearn Random Forest classifiers in Python*. DataCamp Community. Retrieved November 2, 2021, from https://www.datacamp.com/community/tutorials/random-forests-classifier-python.

Paul, S. (2020, January 2). *Feature selection in python sklearn*. DataCamp Community. Retrieved November 2, 2021, from https://www.datacamp.com/community/tutorials/feature-selection-python.

Scikit Learn. (n.d.). *Sklearn.feature_selection.RFE*. scikit. Retrieved November 2, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html#sklearn.feature_selection.RFE.

Scikit Learn. (n.d.). *Sklearn.preprocessing.LabelEncoder*. scikit. Retrieved November 2, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html.