# Capstone Project  1 - Machine Learning

# Factors Influencing ACT Scores

### 1. Unsupervised Learning



### 2. Supervised Learning



### Introduction

In the first section of the machine learning analysis, I approached the problem as an unsupervised learning problem. Using K-Means Clustering method on the average ACT scores from each section (English, Math, Reading, Science), I obtained classes that group the schools based on their ACT scores.

In the second section, I approached the problem as a supervised learning problem. The classes obtained from the first section constitutes the labels of the target . Using the other features in the data such as the percentage of the students who come from different ethnic backgrounds, the percentage of students who receive reduced and free meals, whether a school is public or charter, I try different machine learning algorithms to place the schools into the groups which were created based on the success level.

## 1. Unsupervised Learning

**K-means clustering** algorithm is used to find groups which have not been explicitly labeled in the data. The fundamental step for this unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. In order to find the best number of clusters, I used the elbow method and the Silhouette analysis.

**Elbow method** gives us an idea on what a good k number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids. This method generates a plot of the sum of squared distances for k in a specified range . If the plot looks like an arm, then the elbow on the arm is optimal k. In the graph that I obtained the best number of clusters would be 3 or 4. I also wanted to get a more concrete result using Silhouette analysis.

**Silhouette analysis** is used to determine the degree of separation between clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1]. Silhouette coefficients near +1 indicate that the sample is far away from the neighboring clusters. I applied the Silhouette method to pick the best k among 3, 4 and 5. For k=3, Silhouette coefficient was 0.48, which was the highest among the other scores.

With the k=3 number of clusters, I obtained several plots showing the cluster centroids on different ACT sections. The obtained clusters represented the low, medium and high success levels of schools based on the ACT scores. I also visualized the schools according to their coordinates to see if there was a relation between the location and the success level of the school. But it wasn't very obvious to conclude that there was a specific pattern looking at the schools' locations. Then I created stacked bar graphs to show the proportion of the schools with different success levels in each county. This was more helpful in terms of seeing the counties with more successful or less successful schools.

## 2. Supervised Learning

The aim of this section is to predict the success level of a school based on the percentage of the students who come from different ethnic backgrounds, the percentage of students who receive reduced and free meals, whether a school is public or charter. The first step for **preparing the data** for analysis was converting the values in the categorical column to numerical values. IThen I visualized the distribution of the classes by creating a categorical graph showing the number of schools in each class. Number of schools in each class was not equally distributed. This might cause a problem, because the classifier learns the classes with more samples better and remain weak on the smaller classes. So I used SMOTE technique to balance the number of schools in each class.

**SMOTE** (Synthetic Minority Oversampling Technique) consists of synthesizing elements for the minority class, based on those that already exist. It works randomly picking a point from the minority class and computing the k-nearest neighbors for this point. The synthetic points are added

between the chosen point and its neighbors. In each cluster, there was 2056 data points after applying this technique.

In order to begin applying machine learning algorithms data needs to be split into one train and one split set. In the data that I have, I had information about the schools starting from 2014 and ending in 2018. I split the data into the train set using the information from 2014 to 2017 and the test set using the information only from the year 2018.

**Target and features** are respectively success level of a school, the percentage of American Indian or Alaska Native, Asian, Pacific Islander, Filipino , Hispanic or Latino, African American, White , Two or More Raced students, the percentage of students who receive free or reduced lunch meals and whether a school is public or charter.

Since there are three classes for the machine learning algorithms to predict, I used the algorithms that approaches the problem as a multiclass problem. I applied **Logistic Regression, KNeighbors Classifier** and **Decision Tree Classifier** as a first approach. Then I used ensemble models **Voting, Bagging, Random Forest Classifier** and **XGBoosting** respectively. For most of the models, I used **GridSearchCv** for hyperparameter tuning. For some models, the parameters that I obtained from GridSearchCv resulted in a really high accuracy score on the train set and relatively lower score on the test set which was a sign of overfitting. Then I changed the parameters to be able to reduce the difference of the accuracy scores on the train and test set, but at the same time trying to keep the accuracy as high as possible. For each model, I calculated the accuracy score and obtained the **confusion matrix** and the classification report for a better comparison of the models.

I wanted to determine which feature was most useful for discriminating between the classes to be learned. Based on the Random Forest Classifier, I showed the plot of the feature importances which represented how much including a particular variable improves the prediction. The graph showed that the best predictor of success level of a school is the percentage of the students who receive free and reduced priced meals. The second most important factor is the percentage of Hispanic students. It's followed by the percentage of White and Asian students. Whether the school is public or charter has no effect on the success level of a school.

As a result, the accuracy scores on the test sets are around 75-78%. The best performing models were Voting and XGBoost. Their accuracy scores on test set are the highest and the difference between the accuracy scores on train and test set is small. Future implementations of the models can be developed removing the variables that have no importance and the performance will not suffer.