Capstone_Final_Report

Per the data collected by our organization, we theorized that the populations most affected by school drop-out rates around the world are in less developed regions, are girls, and part of the rural population. To support and substantiate this claim, we used six features or columns of our data frame with the first feature indicating which class out of three each observation or row would fall under, and the other five features containing numerical data of the drop-out rates to see how these features would affect the classification of each observation. To pull this off, we used a Support Vector Machine or SVM to analyze the data and pinpoint the data's class. In our case since we are using a linear classifier for a multiclass or non-linear problem, we employed the kernel parameter within the model. I split the data into both train and test sets and with a test size set to 20% of the entire dataset to permit the SVM adequate data to train on while attempting to give the SVM sufficient data to test on to hone its precision. As can be seen in the confusion matrix mini-diagram of our original prediction and test sets, both the least and less developed regions were correctly predicted more than not based on features utilized (as an aside, many entries in the more developed countries contained null values and for both simplification and SVM requirements to be met, those observations were dropped causing a lack of sufficient training and testing data for this class and therefore receiving erroneous predictions). Based on the model's performance metrics, the accuracy score sits at a satisfactory level to lend support to our theory that the features selected which address the main factors of our problem identification are indeed confirmed by the SVM to fit the criteria for what we believe are the most affected areas and demographics for school drop-out rates. I went ahead further and tested for the efficacy of the SVM on three additional hypothetical datasets; in actuality, just dividing up our original dataset into their original sets of primary, lower secondary, and upper secondary school drop-out rates. For all three datasets, the efficacy of the SVM correctly predicting each observation significantly improved with an accuracy score reaching in the mid seventy percent range. In conjunction with the SVM experiment runs, we've created

three data visualizations to illustrate which areas and populations are more impacted by the drop-out phenomenon. Our first chart which is a bar plot, shows the exponential increase of the average drop-out rate in least developed regions which is over 30%. The other two charts are scatterplots with the first plot highlighting the relationship between male and female drop-out rates. Although the male and female features are technically independent variables, I used the male feature as the independent variable here and the female variable as the dependent variable to contrast and depict the discrepancy of the drop-out rates. Typically, the female drop-out rates outpace the male drop-out rates, but this is clearly not a universal given as illustrated on the plot. I also added a third variable, the 'development regions' to provide further insight as to where the male and female drop-out rates are more prevalent; I input the third variable by using the hue parameter of the scatterplot and more clearly delineated the values of the third variable with the style parameter. Once more, we employed a scatterplot to depict the relationship between the 'urban residence' and 'rural residence' school drop-out rates using the same reasoning and framework as the prior scatterplot and found that on average rural populations suffered more from drop-out rates than urban populations. As per our framework, we used the third variable 'development regions' to validate our theory that the less developed a region is, the higher the drop-out rate impacting rural areas more so.

A few ideas to consider enhancing our Support Vector Model for properly determining our classes or to broaden the scope of coverage for identifying the most disenfranchised regions and populations are latent within the data. To lend support to the data separate from the original set that was used in our original multiclass problem, the development regions feature could be swapped out for the region or sub-region features; as a caveat, this would not work for the countries and areas feature as all the values are unique in that column and would not be applicable to SVM usage. But building on the idea of using either the 'region' or 'sub-region' feature to substitute for class identification, the quintile features which further segment the population according to wealth brackets are features with numerical

values that can either supplement our original dataset used for the SVM or work as a parallel dataset to feed the SVM with to then read the accuracy and supporting metrics to assess the proper classification of that set. Of course, the various features can be tampered around with and mixed to various degrees provided that we reserve one feature which functions as our label or class feature and the other features contain numerical values whether that be integer or rational numbers.

Likewise, modifying our features to input into the SVM in this manner will help us in identifying the regions and populations to prioritize and emphasize UNICEF's social mission in targeting areas that are in need according to urgency of UNICEF's aid. However, it is incumbent upon the user of the SVM to keep in mind that it is a supervised learning algorithm and therefore needs not only the numerical data in place but the accompanying labels as well to decide which factors fall under which labels and thus make up that composite whole or identification. Running and experimenting with different features and labels will help us to tailor our method and outreach efforts if the label classification were used on the 'region' feature for example and dependent on what numerical features were used and demonstrated to have more profound influences on that region's drop-out rate, would guide us in how to customize our social mission in accordance to that region's needs.

Some recommendations for consideration based on our findings of implementing the SVM's classifier and illustrating by way of scatterplots the features accounted for in the SVM, is to focus our social welfare efforts on certain demographics or areas sequentially or to figure in related features to address compositely. To expound on this more, as per the male and female scatterplot, we see that females are more likely to drop-out versus males and with the 'development regions' feature added as a third feature to divulge greater dimensional insight, and after further assessing the particular nation we are at within a region, we can adjust our plan of action accordingly to put precedence on the more disenfranchised demographic and proceed to other demographics from there in offering social welfare assistance programs and services. As a precautionary note, this would be dependent upon thoroughly

gauging if there is a more critical needs demographic or not. Conversely, if we were to factor in and juxtapose an additional scatterplot illustrating different but relevant features to our inquiry, we can decide if there are multiple segments of the population that are adversely affected and need to focus our outreach to deal with them simultaneously. For instance, we could take a two-pronged fork approach and tackle directly both female and rural populations or change our lens in which we are viewing the problem and fine-tune our outreach efforts to target females in rural segments of the population. One can easily change and adjust their target demographic/s and child welfare relief efforts as need be.