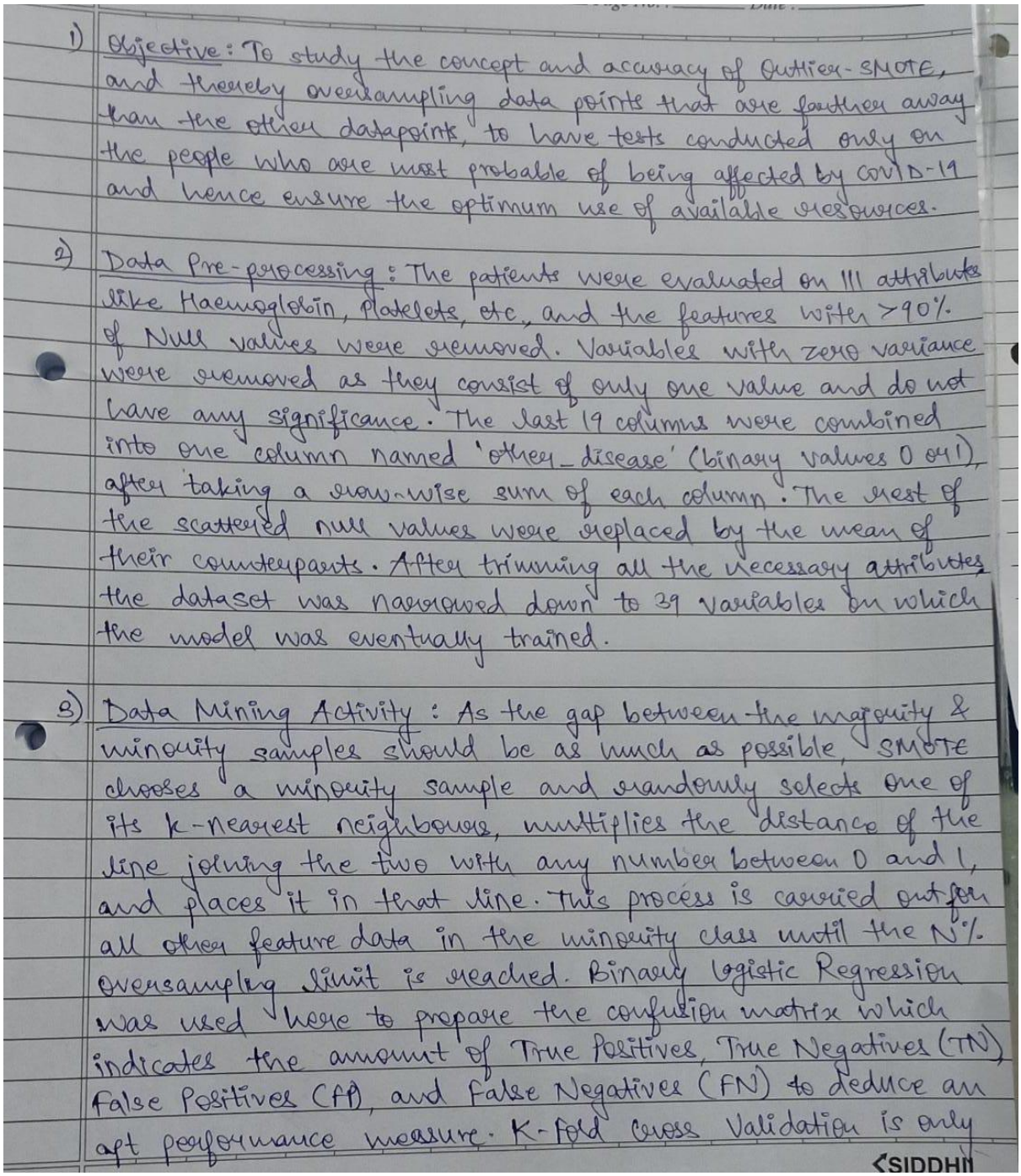
BITS ID – 2020CT04541

Topic: Outlier-SMOTE



1) Objective: To study the concept and accuracy of Outlier-SMOTE, and thereby oversampling data points that are farther away than the other datapoints, to have tests conducted only on the people who are most probable of being affected by COVID-19 and hence ensure the optimum use of available resources.

2) Data Pre-processing: The patients were evaluated on 111 attributes like Haemoglobin, platelets, etc, and the features with >90% of Null values were removed. Variables with zero variance were removed as they consist of only one value and do not have any significance. The last 19 columns were combined into one column named 'other_disease' (binary values 0 or 1), after taking a row-wise sum of each column. The rest of the scattered null values were replaced by the mean of their counterparts. After trimming all the necessary attributes, the dataset was narrowed down to 39 variables on which the model was eventually trained.

3) Data Mining Activity: As the gap between the majority & minority samples should be as much as possible, SMOTE chooses a minority sample and randomly selects one of its k-nearest neighbours, multiplies the distance of the line joining the two with any number between 0 and 1, and places it in that line. This process is carried out for all other feature data in the minority class until the N% oversampling limit is reached. Binary Logistic Regression was used here to prepare the confusion matrix which indicates the amount of True Positives, True Negatives (TN), False Positives (FP), and False Negatives (FN) to deduce an apt performance measure. K-fold Cross Validation is only

SIDDHI

applied on the training data to avoid any data leakage. This process ensures the correct result in any dataset that has tested the algorithm.

4) Metric: This algorithm calculates the amount of times each minority datapoint has to be oversampled. Each dataset has a unique oversampling rate which will help the classifier to achieve maximum accuracy. Here, a fixed oversampling rate of 100% - 500% was considered. After rigorously testing the algorithm on 5-benchmark datasets and comparing with 2 oversampling algorithms, OUTLIER-SMOTE also surpassed in almost every parameter, when applied to the COVID-19 symptoms dataset.

5) Visualization:

→ Step 1: Cleaned dataset is fetched, and is split into majority and minority classes.

→ Step 2: Both classes are split into 90% and 10% of their respective sets.

→ Step 3: There will be 4 classes:
   i) 90% + 10% of the majority set; and
   ii) 90% + 10% of the minority set.

→ Step 4: Combine the 90% of the majority set from step 3(i) and the 90% of the minority set from step 3(ii) and use it as a training set. Similarly, club 10% majority from step 3(i) and 10% majority from step 3(ii) and make it a validation set.

→ Step 5: Train on the decided classifier obtained and store the Recall, Precision, and F1-score.