Data Measurement Strategies

1. **Performance Metrics**:
    - To measure the performance of the models, we will utilize several common evaluation metrics suitable for imbalanced classification problems, including:
        - Accuracy: Measures the overall correctness of the model's predictions.
        - Precision: Determines the proportion of correctly predicted positive instances out of all instances predicted as positive.
        - Recall: Calculates the proportion of correctly predicted positive instances out of all actual positive instances.
        - F1-Score: Harmonic mean of precision and recall, providing a balanced measure between them.
        - Area Under the Receiver Operating Characteristic Curve (AUC-ROC): Evaluates the model's ability to discriminate between positive and negative instances across various classification thresholds.
2. **Confusion Matrix Analysis**:
    - Alongside the performance metrics, we will employ confusion matrix analysis to gain a deeper understanding of the model's predictions. The confusion matrix allows us to examine the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) instances, facilitating a comprehensive assessment of the model's performance.
3. **Statistical Significance Testing**:
    - To assure the reliability of the findings, we will conduct statistical tests of significance. We might use statistical analyses such as t-tests or paired t-tests to determine if the observed differences in performance metrics between models or techniques are statistically significant.

By employing these measurement strategies, we aim to gain insights into the effectiveness of various data balancing techniques on model performance in imbalanced classification tasks. These strategies will enable us to evaluate and compare different techniques objectively and draw meaningful conclusions regarding their impact.

In the next phase of the research project, we will proceed with data preprocessing, model training, and evaluation, applying the described measurement strategies to assess the performance of different models and data balancing techniques.

Please note that the specific implementation details and methodologies may be further refined as the project progresses, ensuring the most accurate and appropriate measurement strategies are employed.

Statistical Analysis

The Wine dataset, obtained from the UCI Machine Learning Repository, has been utilized as a preliminary dataset for this research project. Two datasets are included, related to red and white vinho verde wine samples, from the north of Portugal. The goal is to model wine quality based on physicochemical tests [Cortez et al., 2009].

In this project I concatenated two datasets and formed one dataset and in order to model this problem as an imbalanced classification problem following target definition is used:

- quality <=4, class 0
- quality >4, class 1

We can see target class distribution from Figure 1. This dataset is an example for a moderate imbalance dataset (ratio 1:24)

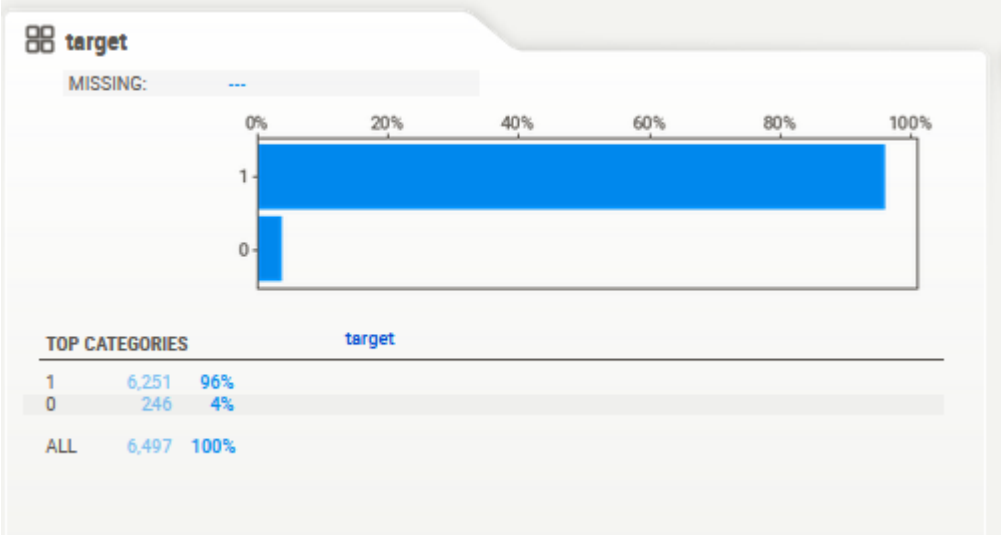

Figure 1. Target Distribution

The Wine dataset is widely recognized and frequently employed in the machine learning community, serving as a standard benchmark dataset for classification tasks.

Note that in this phase of the research project, we utilized a single dataset to demonstrate the measurement strategies. However, it is important to note that in the subsequent stages of the research, multiple datasets of varying degrees of imbalance, including mild, moderate, and extreme imbalances, will be incorporated.

There are 11 features in the dataset and all of them are numerical. In below table you can find basic statistics of these features

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000
mean	7.215307	0.339666	0.318633	5.443235	0.056034	30.525319	115.744574	0.994697	3.218501	0.531268	10.491801
std	1.296434	0.164636	0.145318	4.757804	0.035034	17.749400	56.521855	0.002999	0.160787	0.148806	1.192712
min	3.800000	0.080000	0.000000	0.600000	0.009000	1.000000	6.000000	0.987110	2.720000	0.220000	8.000000
25%	6.400000	0.230000	0.250000	1.800000	0.038000	17.000000	77.000000	0.992340	3.110000	0.430000	9.500000
50%	7.000000	0.290000	0.310000	3.000000	0.047000	29.000000	118.000000	0.994890	3.210000	0.510000	10.300000
75%	7.700000	0.400000	0.390000	8.100000	0.065000	41.000000	156.000000	0.996990	3.320000	0.600000	11.300000
max	15.900000	1.580000	1.660000	65.800000	0.611000	289.000000	440.000000	1.038980	4.010000	2.000000	14.900000

Table 1. Basic Statistics of features

Now let's use statistical tests in order to see which variables are significant for predicting our target variable. Since all our features are numerical and our target is binary categorical data, we can use independent samples t-test for the tests. Below figure shows the results:

```
Column:fixed acidity
t:-1.7568
p_value:0.079
==========================
Column:volatile acidity
t:-12.3295
p_value:0.0
==========================
Column:citric acid
t:4.9892
p_value:0.0
==========================
Column:residual sugar
t:3.934
p_value:0.0001
==========================
Column:chlorides
t:-2.782
p_value:0.0054
==========================
Column:free sulfur dioxide
t:6.8918
p_value:0.0
==========================
Column:total sulfur dioxide
t:2.8428
p_value:0.0045
==========================
Column:density
t:-1.318
p_value:0.1876
==========================
Column:pH
t:-1.6208
p_value:0.1051
==========================
Column:sulphates
t:2.7454
p_value:0.0061
==========================
Column:alcohol
t:4.1269
p_value:0.0
==========================
```

Figure 2. t-test results

As we can see there is not significant relationship between variables "fixed acidity", "chlorides", "density", "pH" and target variable. In the next part of this report (Data Visualisation) we will confirm these findings with the help of boxplot analysis.

Data Visualizations

Let's first look at associations between variables. In Figure 3 we can see correlations between features and target. There is no highly correlated variable with target variables, which can cause to poor model performance. In case of that we can consider changing the dataset. Although the graph doesn't show exact values for correlations it is a nice way to explore associations.

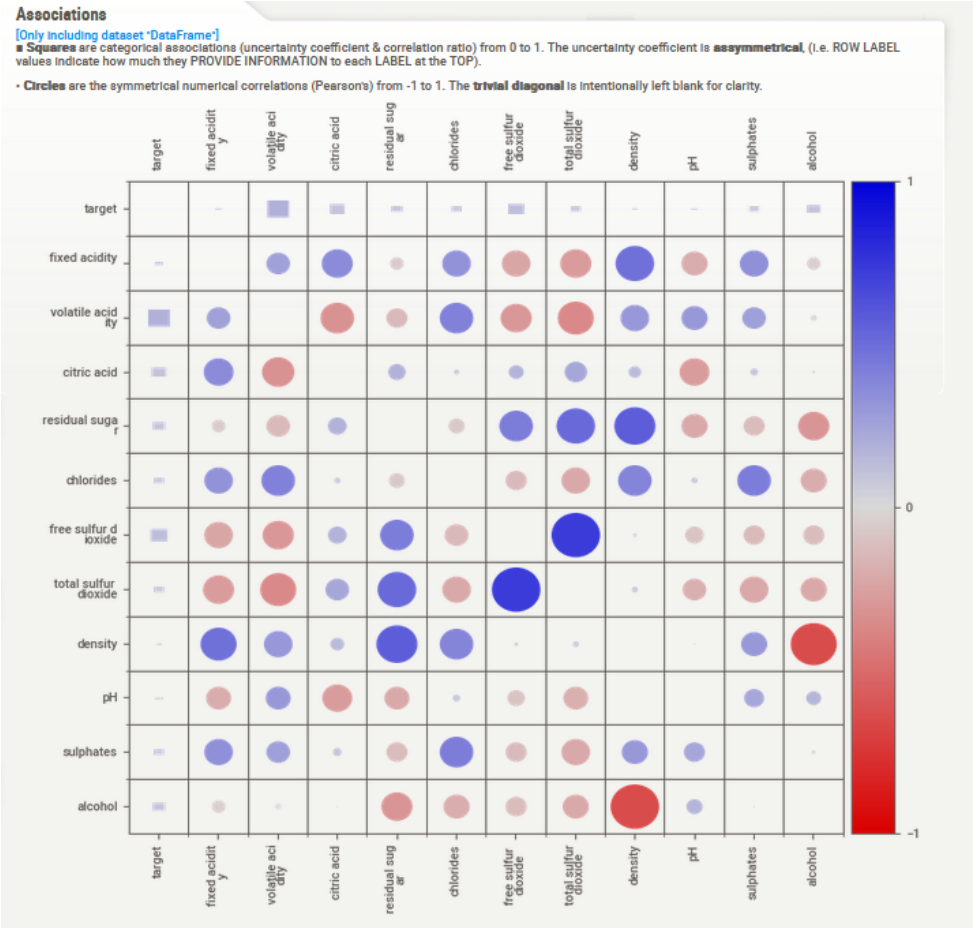

Figure 3. Associations between variables

Boxplot Analysis

Let's do some boxplot analysis. As we talked in previous section there is no relationship between "fixed acidity" and target. We can see that clearly in Figure 4.

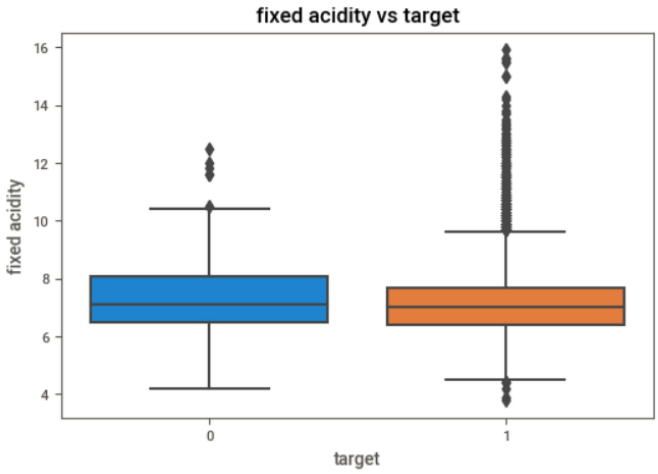

Figure 4. Boxplot of “fixed acidity” and target

In Figure 5 we clearly see that there is a difference between means of two classes. Therefore we can say that there is relationship between “volatile acidity” and target

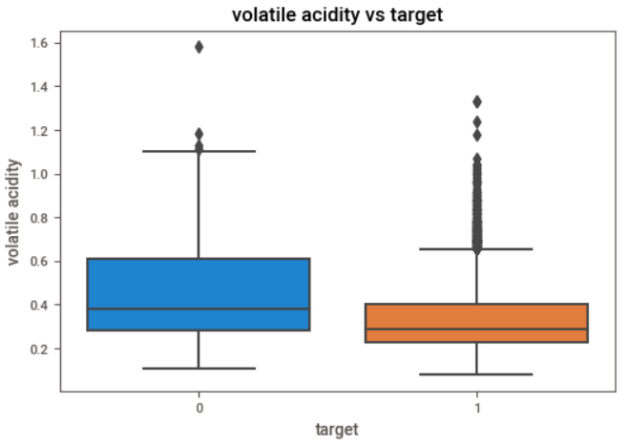

Figure 4. Boxplot of “volatile acidity” and target

Conclusion

In this report, we have discussed the progress made in the research project titled "Exploring Class Imbalance Solutions: Investigating the Effectiveness of Data Balancing Techniques on Model Performance." The report focused on three key areas: the description of measurement strategies, statistical analysis, and visualization of the data.

Firstly, we outlined the measurement strategies employed in the research project. These strategies encompassed the selection of performance metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. These measurement strategies will enable us to draw meaningful conclusions regarding the effectiveness of different data balancing techniques.

Secondly, we conducted a statistical analysis similar to our in-class work. Through statistical significance testing, we will determine whether observed differences in performance metrics between models or techniques are statistically significant. This analysis will add rigor to our findings and enhance the credibility of our research outcomes.

Lastly, we highlighted the importance of data visualization. Visual storytelling plays a vital role in conveying complex information in a clear and intuitive manner.

In conclusion, this report marks the initial steps of our research project, providing an overview of the measurement strategies, statistical analysis, and the upcoming visual storytelling component. The next phase of the research will focus on implementing the proposed strategies, analyzing the results, and further refining our understanding of the research problem.

As we progress further, we anticipate encountering certain risks and challenges, such as dataset limitations, the selection of appropriate data balancing techniques, and potential trade-offs between model performance and computational efficiency. However, with careful planning, thorough analysis, and continuous evaluation, we are confident in overcoming these challenges and producing valuable findings that contribute to the existing body of knowledge in the field of imbalanced classification.

Moving forward, we will continue to execute the research plan, collect and analyze data, refine our measurement strategies, and interpret the results. Through our dedication and rigorous approach, we aim to provide actionable insights and practical recommendations to address the class imbalance problem and improve the performance of classification models.