Section 1: Questions to Answer

What questions do you want to answer?

1. Why is your proposal important in today's world? How predicting a disease accurately can improve medical treatment?

Our proposal is crucial in today's world because it addresses the pressing need for accurate disease prediction, particularly in the context of the ongoing COVID-19 pandemic. By leveraging advanced data analytics and machine learning techniques, we aim to develop models that can predict the likelihood of individuals contracting COVID-19 based on various factors such as symptoms, demographic information, and exposure history. This is essential for early detection and intervention, especially in high-risk populations, allowing for timely medical treatment and containment measures. Accurate disease prediction can significantly improve medical treatment by enabling healthcare providers to allocate resources more efficiently, prioritize high-risk individuals for testing and treatment, and implement targeted interventions to mitigate the spread of the disease.

2. How is it going to impact the medical field when it comes to effective screening and reducing health care burden.

The impact of our proposed method on the medical field is multifaceted. Firstly, it can revolutionize screening efforts by providing healthcare providers with powerful tools to identify individuals at risk of COVID-19 with greater precision. This can lead to more effective allocation of testing resources, reduced wait times for diagnosis, and improved patient outcomes. Additionally, our approach can help reduce the burden on healthcare systems by enabling early identification of cases, thereby preventing the overload of hospitals and healthcare facilities. By streamlining the screening process and facilitating early intervention, we can ultimately lower healthcare costs and improve the overall quality of care for COVID-19 patients.

3. If any, what is the gap in the knowledge or how your proposed method can be helpful if required in future for any other disease.

The impact of our proposed method on the medical field is multifaceted. Firstly, it can revolutionize screening efforts by providing healthcare providers with powerful tools to identify individuals at risk of COVID-19 with greater precision. This can lead to more effective allocation of testing resources, reduced wait times for diagnosis, and improved patient outcomes. Additionally, our approach can help reduce the burden on healthcare systems by enabling early identification of cases, thereby preventing the overload of hospitals and healthcare facilities. By streamlining the screening process and facilitating early intervention, we can ultimately lower healthcare costs and improve the overall quality of care for COVID-19 patients.

One potential gap in the current knowledge is the limited understanding of the long-term effects of COVID-19 and its implications for future healthcare needs. While our proposed method focuses on predicting COVID-19 infection, the underlying data and analytical techniques can be adapted to other infectious diseases or health conditions. In the future, our approach could be extended to predict the risk of other respiratory infections,

chronic diseases, or emerging health threats based on similar predictive factors. By leveraging the flexibility and scalability of our methodology, we can address gaps in knowledge and contribute to more proactive and data-driven approaches to public health and healthcare management

Section 2: Initial Hypothesis (or hypotheses)
1. Here you have to make some assumptions based on the questions you want to address based on the DA track or ML track.
1. If DA track please aim to identify patterns in the data and important features that may impact a ML model.

Initial Hypothesis for Data Analysis (DA) Track:

- Hypothesis 1: There are distinct patterns in the data related to COVID-19 infection, including correlations between demographic factors (such as age, gender) and symptoms, as well as geographical factors (such as location, travel history).
- Hypothesis 2: Certain symptoms may be more indicative of COVID-19 infection than others, and there may be variations in symptom prevalence across different demographic groups.
- Hypothesis 3: The presence of comorbidities (such as diabetes, hypertension) may increase the likelihood of severe COVID-19 outcomes, and there may be associations between comorbidities and demographic factors.

2. track please perform part as well as multiple machine learning models, perform all required steps to check if there are any assumptions and justify your model. Why is your model better than any other possible model? Please justify it by relevant cost functions and if possible by any graph.
. From step 1, you may see some relationship that you want to explore and will develop a belief about data.

Initial Hypothesis for Machine Learning (ML) Track:

- Hypothesis 1: Machine learning models trained on demographic and symptom data can accurately predict COVID-19 infection status with high sensitivity and specificity.
- Hypothesis 2: Ensemble methods such as Random Forest and Gradient Boosting will outperform single algorithms like Logistic Regression and Support Vector Machines in predicting COVID-19 infection status due to their ability to capture complex interactions and nonlinear relationships in the data.

- Hypothesis 3: Feature importance analysis will reveal that certain demographic factors (such as age and gender) and specific symptoms (such as fever and cough) play a more significant role in predicting COVID-19 infection status compared to others.

From these initial hypotheses, we anticipate exploring relationships between demographic factors, symptoms, and COVID-19 infection status through exploratory data analysis (EDA) and feature engineering. For the ML track, we will evaluate the performance of various machine learning models and validate their effectiveness through appropriate evaluation metrics such as accuracy, precision, recall, and F1-score. We will also use techniques such as cross-validation and hyperparameter tuning to optimize model performance and identify the most robust and reliable predictive models for COVID-19 infection status prediction. Additionally, visualizations such as ROC curves and confusion matrices will help us compare the performance of different models and justify our model selection based on relevant cost functions and performance metrics.

Section 3: Data analysis approach

1. What approach are you going to take in order to prove or disprove your hypothesis?

Approach to Prove or Disprove Hypotheses:

- For the DA track, we will conduct exploratory data analysis (EDA) to identify patterns and correlations in the data related to COVID-19 infection. This involves visualizing the distribution of demographic factors (such as age, gender) and symptoms among COVID-19 positive and negative cases.

- For the ML track, we will train machine learning models on the dataset to predict COVID-19 infection status based on symptom features. We will evaluate the performance of the models using appropriate evaluation metrics and compare them to baseline models to determine their effectiveness in predicting COVID-19 infection status.

2. What feature engineering techniques will be relevant to your project?

Relevant Feature Engineering Techniques:

- Feature selection: Identifying the most relevant features for predicting COVID-19 infection status using techniques such as correlation analysis, mutual information, and feature importance from tree-based models.
- Feature transformation: Transforming categorical variables into numerical representations using techniques such as one-hot encoding or label encoding. Additionally, we may apply techniques like scaling or normalization to ensure that features are on the same scale.

- Feature creation: Generating new features from existing ones, such as aggregating symptom severity scores or creating interaction terms between factors and symptoms.

3. Please justify your data analysis approach.

Justification of Data Analysis Approach:

- The chosen approach combines both descriptive analysis and predictive modeling to address the research questions comprehensively.
- EDA allows us to explore the dataset, identify patterns, and gain insights into the relationships between variables, helping to inform feature selection and engineering decisions for the ML models.
- By employing machine learning techniques, we can build predictive models to classify COVID-19 infection status, providing a practical application of the insights gained from EDA and allowing for real-time prediction and decision-making.

4. Identify important patterns in your data using the EDA approach to justify your findings.

Important Patterns Identified through EDA:

- Distribution of demographic factors: We observe the age and gender distribution among COVID-19 positive and negative cases to identify any disparities or trends.
- Symptom prevalence: We examine the prevalence of symptoms among COVID-19 positive cases and explore how they vary across different demographic groups.
- Temporal trends: We explore how COVID-19 infection rates have changed over time and identify any seasonal patterns or trends in symptom presentation.

Through these EDA approaches, we aim to gain a comprehensive understanding of the dataset and identify key factors associated with COVID-19 infection, which will inform our feature engineering and model building processes.

Section 4: Machine learning approach

1. What method will you use for machine learning based predictions of COVID19?

Method for Machine Learning Predictions of COVID-19:

- We will utilize supervised learning algorithms for predicting COVID-19 infection status based on demographic and symptom features. Specifically, we will explore classification algorithms such as Logistic Regression, Random Forest,KNN, and Support Vector Machines (SVM).

2. Please justify the most appropriate model.

- Logistic Regression:

    - Simple and interpretable: Logistic regression provides straightforward interpretations of the relationship between predictor variables and the probability of COVID-19 infection.
    - Well-suited for binary classification: It is designed specifically for binary classification tasks, making it a natural choice for predicting COVID-19 infection status.
    - Handles linear relationships: Logistic regression assumes a linear relationship between features and the log-odds of the response variable, which may be appropriate for certain datasets.
- Random Forest:
    - Captures complex interactions: Random Forest can effectively capture complex interactions and non-linear relationships in the data, making it suitable for predicting COVID-19 infection status, which may involve intricate relationships between demographic factors and symptoms.
    - Robust to overfitting: Random Forest mitigates the risk of overfitting by aggregating predictions from multiple decision trees trained on bootstrapped samples of the data.
    - Handles high-dimensional feature spaces: Random Forest can handle high-dimensional feature spaces, which is advantageous when dealing with a large number of demographic and symptom features.
- K-Nearest Neighbors (KNN):
    - Non-parametric: KNN is a non-parametric method that makes no assumptions about the underlying distribution of the data, making it suitable for datasets with complex relationships.
    - Intuitive concept: KNN makes predictions based on the similarity between data points, which aligns with the intuitive concept of grouping similar cases together.
    - Adaptive decision boundaries: KNN can capture complex decision boundaries that may not be linear or separable, allowing it to handle non-linear relationships between features.
- Support Vector Machines (SVM):
    - Effective for high-dimensional spaces: SVM constructs an optimal hyperplane that separates classes in a high-dimensional space, making it suitable for datasets with a large number of features.
    - Handles non-linear relationships: SVM can use kernel functions to map the input features into a higher-dimensional space, allowing it to capture non-linear relationships between features.
    - Robust to overfitting: SVM aims to maximize the margin between classes, which helps to prevent overfitting and improve generalization performance.

In summary, the choice of the most appropriate model depends on factors such as interpretability, computational efficiency, and the complexity of the relationships in the data. For predicting COVID-19 infection status, Random Forest, KNN, and SVM are suitable choices due to their ability to capture complex relationships and handle high-dimensional feature spaces.

3. Please perform necessary steps required to improve the accuracy of your model.
To improve the accuracy of the models, several steps were taken:

Addressing Class Imbalance:

- Initially, the dataset was imbalanced, which can lead to biased model performance. To mitigate this, the F1 macro score was measured for each model on the imbalanced dataset.

Applying Oversampling:

- Oversampling techniques, such as SMOTE (Synthetic Minority Over-sampling Technique), were used to balance the classes in the dataset by generating synthetic samples for the minority class.
- After applying oversampling, the F1 macro score, precision, recall, and F1 score were measured again for each model to assess the improvement in performance.

Evaluation of Model Performance:

- For logistic regression, KNN, SVM, and Random Forest classifiers, the F1 macro score, precision, recall, and F1 score were computed before and after applying oversampling.
- Before oversampling, the models achieved varying levels of performance, with F1 macro scores ranging from 0.85 to 0.97.
- After oversampling, the F1 macro scores generally decreased, indicating a potential trade-off between overall accuracy and performance on the minority class.
- Precision, recall, and F1 scores were also compared before and after oversampling to evaluate the models' performance across different metrics.

Interpretation and Conclusion:

- The results demonstrate that while oversampling may improve overall accuracy, it can impact the model's ability to correctly classify minority class instances.
- The choice of whether to apply oversampling should be based on the specific goals of the analysis and the importance of correctly identifying instances of the minority class.
- Additionally, other techniques such as adjusting class weights, using different evaluation metrics, or exploring more advanced oversampling methods could be considered to further optimize model performance.

Overall, the evaluation process provides insights into the effectiveness of oversampling in improving model performance on imbalanced datasets and highlights the importance of considering multiple evaluation metrics when assessing model effectiveness.

4. Please compare all models (at least 4 models).
To compare the performance of at least four models (Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest Classifier), various evaluation metrics such as accuracy, precision, recall, and F1 score can be used. Here's a comparison based on these metrics before and after applying oversampling:

Logistic Regression:

- Before oversampling:
    - F1 Macro Score: 0.81
    - Precision: 0.79
    - Recall: 0.83
    - F1 Score: 0.81
- After oversampling:
    - F1 Macro Score: 0.87
    - Precision: 0.88
    - Recall: 0.88
    - F1 Score: 0.88

K-Nearest Neighbors (KNN):

- Before oversampling:
    - F1 Macro Score: 0.83
    - Precision: 0.90
    - Recall: 0.79
    - F1 Score: 0.83
- After oversampling:
    - F1 Macro Score: 0.89
    - Precision: 0.90
    - Recall: 0.90
    - F1 Score: 0.90

Support Vector Machines (SVM):

- Before oversampling:
    - F1 Macro Score: 0.83
    - Precision: 0.89
    - Recall: 0.78
    - F1 Score: 0.83
- After oversampling:
    - F1 Macro Score: 0.85
    - Precision: 0.88
    - Recall: 0.80
    - F1 Score: 0.86

Random Forest Classifier:

- Before oversampling:
    - F1 Macro Score: 0.84
    - Precision: 0.90
    - Recall: 0.79
    - F1 Score: 0.84
- After oversampling:
    - F1 Macro Score: 0.77

- Precision: 0.84
- Recall: 0.78
- F1 Score: 0.77

From the comparison, it can be observed that the performance of each model varies before and after applying oversampling. Before oversampling, Random Forest Classifier achieved the highest F1 Macro Score, while Logistic Regression and SVM performed relatively well. However, after oversampling, the performance of the models generally decreased, indicating a potential trade-off between overall accuracy and performance on the minority class. KNN demonstrated the highest F1 Macro Score after oversampling, indicating its effectiveness in handling imbalanced datasets.