COVID-19 Predictions Based on Symptoms

Technological advancements have a profound effect on all spheres of life, whether in the medical field or any other domain. Artificial intelligence has shown promising results in healthcare by making decisions based on the analysis and processing of data. Early diagnosis is crucial in preventing the spread and development of life-threatening diseases. COVID-19, being highly contagious, has become a global epidemic that demands swift action.

The rapid spread of the virus necessitates the development of a system for its detection. With the increasing use of technology, a wealth of data about COVID-19 is readily available, which can be harnessed to gather essential information about the virus. In this project, we compared the accuracy of various machine learning algorithms in predicting COVID-19 outcomes and selected the most accurate one for the final model testing. With just eight binary features, our model achieved high accuracy in predicting COVID-19 test results. This paper suggests a practical solution through the development of a health monitoring system that can mitigate the impact of COVID-19.

Section 1: Questions to Answer

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

Predicting diseases accurately is crucial today because it enables early intervention, personalized treatment, efficient resource allocation, reduced healthcare costs, improved patient outcomes, accelerated drug development, enhanced public health responses, and empowers patients to take control of their health.

Modern healthcare relies heavily on data analysis, machine learning, and artificial intelligence to make accurate disease predictions. These technologies can process vast amounts of patient data, including genetic information, medical history, and lifestyle factors, to generate precise risk assessments. As more data becomes available and these models continue to improve, the potential for accurate disease prediction and personalized treatment will only increase, leading to better healthcare outcomes and a healthier global population.

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

Accurate disease prediction and effective screening have a profound impact on healthcare. They enable early detection, leading to less expensive and invasive treatments, and reducing healthcare costs. Preventive measures based on risk assessments prevent diseases, further decreasing the burden on the healthcare system. Efficient resource allocation ensures timely care, improving patient outcomes. Public health benefits from timely responses to outbreaks. Personalized treatments based on predictions enhance patient satisfaction. Disease prediction also accelerates research and drug development. These advancements create a proactive healthcare model that lowers costs, improves outcomes, and benefits both individuals and healthcare systems.

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

Disease prediction methods often face a knowledge gap due to the necessity of extensive and diverse datasets, which should encompass factors like genetics, medical history, lifestyle, and environmental influences for accurate predictions. This data challenge is particularly notable for rare diseases or underserved populations. Nevertheless, the proposed method demonstrates significant adaptability and potential utility for upcoming diseases or healthcare challenges. Inwhile initially focused on specific diseases, it serves as a foundational and flexible approach to disease prediction, capable of refinement, expansion, and integration with new data sources to effectively tackle future healthcare complexities, establishing itself as a valuable tool in the ever-evolving landscape of medical research and practice.

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.**
    - **If DA track, please aim to identify patterns in the data and important features that may impact a ML model.**
    - **If ML track please perform part 1 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.**

In my task, I need to make assumptions based on the questions I want to address in either the Data Analysis (DA) track or the Machine Learning (ML) track. Here's what I should aim for in each track:

- For the Data Analysis (DA) Track: identify patterns and important features.

  1- Data Exploration: By exploring my dataset to understand its structure, summary statistics, and distribution of variables.

  2- Feature Analysis: Identify and analyze the key features or variables in your dataset. By using feature importance techniques to assess the impact of features on target variables.

  3- Data Visualization: Create data visualizations (e.g., heatmaps, histograms) to visualize relationships between variables and discover any patterns.

  4- Feature Engineering: Perform feature engineering, including handling missing data, encoding categorical variables using different encoding methods, and removing unnecessary features.

- For the Machine Learning (ML) Track: Start by performing data preprocessing, exploratory data analysis, and feature engineering to better understand your dataset. Afterward, you will proceed to build multiple machine-learning models to predict COVID-19-positive cases.

  1-Model Selection: Choose appropriate machine learning models based on the nature of my data.

  2-Split Data: Split my dataset into training and testing sets for model evaluation.

  3-Model Training: Train multiple machine learning models (e.g. Logistic Regression, Decision Trees, Random Forest, Support vector Machine, KNN) on my training data.

  4-Hyperparameter Tuning: Optimize model hyperparameters

  5-Model Evaluation: Evaluate model performance using relevant metrics (accuracy, precision, recall, F1-score, etc.).

  6-Visualizations: Create visualizations (e.g. Confusion matrices) to explain my model behaviors.

  7-Model Comparison: Compare the performance of different models and justify the best model based on confusion metrics.

  In both tracks, the key is to use your domain knowledge and initial data exploration to formulate hypotheses and make informed decisions throughout the data analysis or machine learning process. Your goal is to provide a clear rationale for the choices you make and justify why you believe your approach is suitable for addressing the questions or problems you're aiming to solve.

Section 3: Data analysis approach

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

To prove or disprove the initial hypotheses related to predicting COVID-19 diagnosis based on symptoms, the following data analysis approach will be taken:

1-Exploratory Data Analysis (EDA): By using EDA, gain insights into the dataset and uncover any initial patterns or relationships between variables. This will include:

- Data Visualization: Create various visualizations such as histograms, bar charts, Pie charts and heatmaps to visualize the distribution of features, and relationships among variables, and identify any trends.
- Summary Statistics: Calculate summary statistics (mean, median, mode, standard deviation, etc.) to understand the central tendencies and variability of different features.

2-Feature Engineering: Perform feature engineering to prepare the dataset for machine learning model training. This includes:

- Handling Missing Data: Identify and handle missing values in the dataset through imputation or removal of rows/columns with missing data.
- Encoding Categorical Variables: Encode categorical variables into numerical format suitable for machine learning models.

3-Feature Importance Analysis: Determine the importance of each feature with respect to COVID-19 diagnosis

4-Model Validation: Evaluate machine learning models using appropriate evaluation metrics (accuracy, precision, recall, F1-score,) to assess their predictive performance.

This validation step helps in validating or disproving the hypotheses by quantifying the model's ability to predict COVID-19 diagnosis accurately.

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

Feature engineering is the pre-processing step of machine learning, which is used to transform raw data into features that can be used for creating a predictive model using Machine learning. In the project I've been described, which involves predicting COVID-19 test results, several feature engineering techniques can be relevant:

- One-Hot Encoding: this technique to convert categorical variables into a numerical format.
- Label Encoding: label encoding to convert binary categorical features into numeric format.
- Drop features.
- Handling Missing Values: Handling missing values using techniques like forward fill and backward fill. other methods such as mean imputation, median imputation, and mode imputation.
- Imbalanced Data: Handling class imbalance through under-sampling.

The choice of feature engineering techniques depends on the specific dataset, the algorithms that I plan to use, and the problem I am trying to solve. It's often an iterative process where I try different techniques and evaluate their impact on model performance.

3. Please justify your data analysis approach.

The data analysis approach which involves predicting COVID-19 test results is:

1. Exploratory Data Analysis (EDA): EDA is essential to gain a deep understanding of the dataset.
2. Data Preprocessing and Cleaning: Preprocessing and cleaning the data is the first step in the data analysis project.
3. Feature Engineering: Feature engineering is a critical step in predictive modelling. This process is used to transform and create features that can capture relevant information for predicting COVID-19 test results. Techniques like one-hot encoding, label encoding, and handling missing values are used to prepare the data for modelling.

4. Dealing with Imbalanced Data: Addressing class imbalance is crucial in medical diagnostics, where positive cases are often much less frequent than negative cases. Under-sampling is one way to tackle this issue, and it's justified because it helps ensure that the model is not biased towards the majority class.
5. Model Selection and Evaluation: The choice of models and evaluation metrics depends on the problem.
6. Hyperparameter Tuning: Hyperparameter tuning ensures that the chosen models perform optimally. It involves adjusting model parameters like learning rates, tree depths, criterion, and estimators.
7. validation and Testing: The final model is validated on a separate test dataset to assess its real-world performance.
8. Monitoring and Updating: Finally, in a real-world medical application, continuous monitoring and updating of the model are justified. COVID-19 trends, diagnostic criteria, and testing protocols may change over time, and the model should be adaptable to these changes.

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

Exploratory Data Analysis refers to the crucial process of performing initial investigations on data to discover patterns to check assumptions with the help of summary statistics and graphical representations.

EDA gives an idea about missing values in the dataset.

EDA helps in identifying meaningful patterns in data. In my dataset, the target column is imbalanced.

There are 5 unique values in the 'Cough symptoms', 'Fever', 'Sore throat', 'Shortness_of_breath', and 'Headache' columns, and 3 unique values in the 'Age_60_above', 'Sex', and 'Known_contact' columns in the input data. These findings were identified during the EDA.

Section 4: Machine learning approach

1. What method will you use for machine learning based predictions of COVID-19?

To use machine learning for predictions on specific datasets related to COVID-19, I would typically follow these steps:

- Data Collection: Collect dataset. Ensure that it includes relevant features and labels for the prediction task.
- Data Preprocessing: Prepare data for machine learning. This involves tasks like handling missing values, encoding categorical variables, and splitting the data into training and testing sets.
- Feature Engineering: Depending on the nature of the data, I must transform my data set to make it more informative for the model.
- Model Selection: Choose an appropriate machine learning model based on the problem that I must solve.
- Model Training: Train my model on the training dataset. Tune hyperparameters if necessary to optimize model performance.
- Model Evaluation: Evaluate the model's performance on a separate test dataset using appropriate evaluation metrics. Common metrics include accuracy, precision, recall, F1-score, etc.
- Iterate and Refine: Based on the evaluation results, I have to iterate on my model, make adjustments, and retrain it to improve its performance.
- Deployment: Once you are satisfied with your model's performance, you can deploy it to make predictions on new, unseen data.

2-Please justify the most appropriate model.

In the context of a COVID-19 dataset where the goal is to predict whether a patient is positive or negative for COVID-19 based on test data, choosing the most appropriate model involves considering the evaluation metric, which in this case is recall. Recall is an important metric in situations where identifying true positives is crucial, even if it means having some false positives.

- Based on the recall values
- Logistic Regression (LR): 74.54
- Decision Tree: 76.51
- Random Forest: 76.40
- Support Vector Machine (SVM): 76.40
- K-Nearest Neighbours (KNN): 78.24

It appears that K-Nearest Neighbours (KNN) has the highest recall value (78.24), which means it is the best at correctly identifying COVID-19-positive patients among the models I've tested.

3.Please perform the necessary steps required to improve the accuracy of your model.

To improve the accuracy of my machine learning model for predicting COVID-19-positive cases, I followed a systematic process that included data preprocessing, feature engineering, model selection, hyperparameter tuning, and performance evaluation. I have applied all the necessary steps to improve the accuracy of the model to my COVID dataset.

4.Please compare all models (at least 4 models).

Table gives the summary of the results of all the models.

SR. No	Model Name	Accuracy	Precision	Recall	F1 Score
1	Logistic Regression	83.53	91.02	74.54	81.95
2	Decision Tree	85.01	92.36	76.51	83.69
3	Random Forest	85.13	92.74	76.4	83.78
4	Support Vector Machines	85.1	92.66	76.4	83.74
5	KNN	84.75	90.11	78.24	83.75

Weighted average of F1 score, precision and recall is used to compare the models since the dataset was imbalanced. From the above summary, it can be concluded that the model with highest accuracy on the covid19 dataset is Random Forest.