Title: AIDI 1002 Final Term Project Report

Group: MLP Project

Members' Names or Individual's Name: Gurleen kaur and Amanpreet kaur

Emails: 200604519@student.georgianc.on.ca and 200564069@student.georgianc.on.ca

Introduction

The project is aimed at improving COVID-19 diagnosis through the comparison of various machine learning models. Through the performance evaluation of each model using measures such as accuracy, ROC score, and mean squared error (MSE), the project aims to establish the best predictive model for COVID-19 diagnosis. The project is a comprehensive review in comparison to previous studies that used Random Forest, K-nearest neighbors, and Naive Bayes algorithms.

Problem Statement and Solution:

The global COVID-19 pandemic highlighted the need for prompt and accurate diagnostic tools to control the virus's spread and manage healthcare resources. Traditional testing methods, like PCR, are often limited by accuracy, cost, and processing time. Machine learning offers a scalable and effective alternative for developing predictive systems. This project addresses the limitation of previous studies, which focused on a narrow range of algorithms, by introducing a wider array of models, including Deep Neural Networks (DNN), Decision Trees, and XGBoost. These models, with their diverse learning capabilities, are used to capture intricate data patterns.

Dataset Analysis:

The project uses a dataset named Covid Dataset.csv containing 5,434 entries and 21 columns. The dataset is composed entirely of categorical features with 'Yes' and 'No' values. The columns represent various symptoms and risk factors.

The target variable is COVID-19, which indicates whether a patient is positive or negative. Initial data exploration revealed no missing values. The columns Wearing Masks and Sanitization from Market were later dropped from the model's feature set.

Methodology:

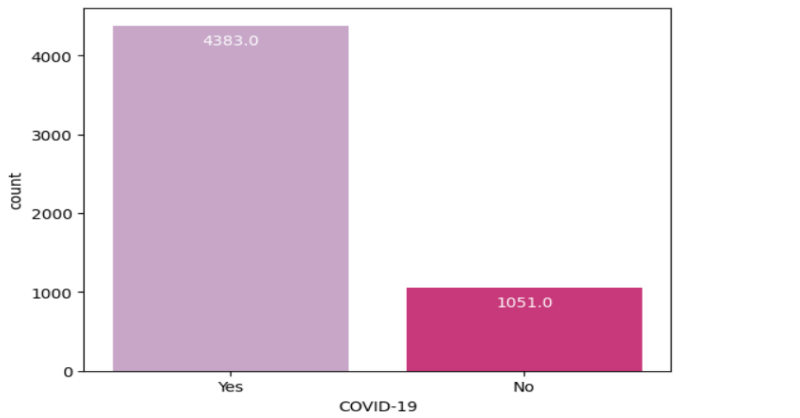
The methodology followed a structured machine learning pipeline:

1. **Data Preprocessing:** The categorical 'Yes' and 'No' values in all columns were transformed into numerical '1' and '0' using LabelEncoder.
2. **Exploratory Data Analysis (EDA):** The distribution of the target variable and its relationship with other features were visualized using bar charts and pie charts.
3. **Data Splitting:** The preprocessed dataset was split into training and testing sets, with 80% of the data used for training and 20% for testing.

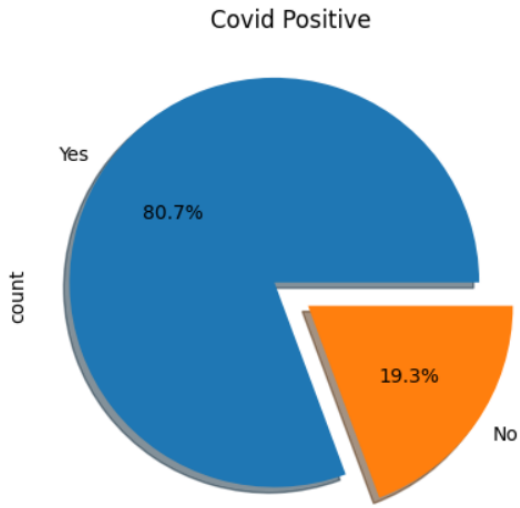
	missing_values	percent_missing %
Breathing Problem	0	0.0
Fever	0	0.0
Dry Cough	0	0.0
Sore throat	0	0.0
Running Nose	0	0.0
Asthma	0	0.0
Chronic Lung Disease	0	0.0
Headache	0	0.0
Heart Disease	0	0.0
Diabetes	0	0.0
Hyper Tension	0	0.0
Fatigue	0	0.0
Gastrointestinal	0	0.0
Abroad travel	0	0.0
Contact with COVID Patient	0	0.0
Attended Large Gathering	0	0.0
Visited Public Exposed Places	0	0.0
Family working in Public Exposed Places	0	0.0
Wearing Masks	0	0.0
Sanitization from Market	0	0.0
COVID-19	0	0.0

Figure shows Histogram of all the attributes

4. **Model Training and Evaluation:** A wide range of machine learning models were trained and evaluated. The performance was assessed using several metrics: accuracy, ROC AUC score, mean squared error (MSE), and R-squared score. The models included:
    - **Previous Models:** Logistic Regression, K-Nearest Neighbors (KNN), Random Forest.
    - **Contributed Models:** Deep Neural Network (DNN), XGBoost, Decision Tree, Gaussian Naive Bayes.
5. **Hyperparameter Tuning:** GridSearchCV was used for hyperparameter tuning on the KNN and Random Forest models to find the optimal parameters and improve performance.



Histogram showing the number of patients with covid positive and negative



Pie chart showing the number of patients with covid positive and negative

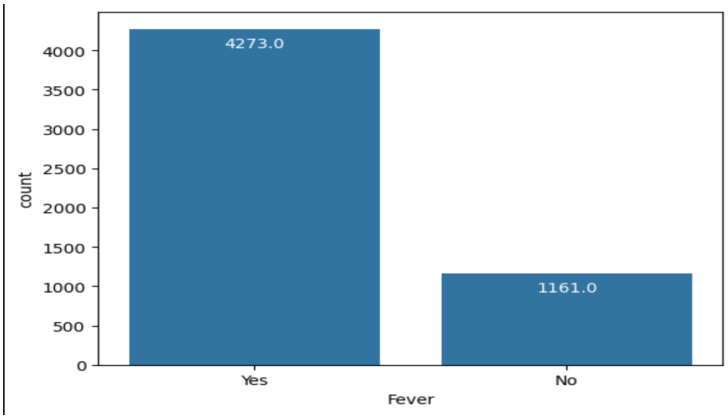
IMPLEMENTATION

With the growth of computer technology,predictive mod-eling is changing.We are now able to make Predictable modeling more effcient, and less expensive than before.In our project,we use various class If cation algorithms to predict and use a gridsearchCV to fnd the most advanced solution for each algorithm. Some of the categorization algorithms that have been employed include:

LogisticRegression

Logistic regression is a data categorization technique that uses machine learning.This algorithm, Models the odds of the potential outcomes of a single experiment using alogistic function.The easiest way

To understand the influence of numerous independent factors on a single outcome variable is to use logistic regression, which was designed for this purpose. In general,the algorithm calculates the probability of belonging to a particular class.We have two classes here,
y=0,1.

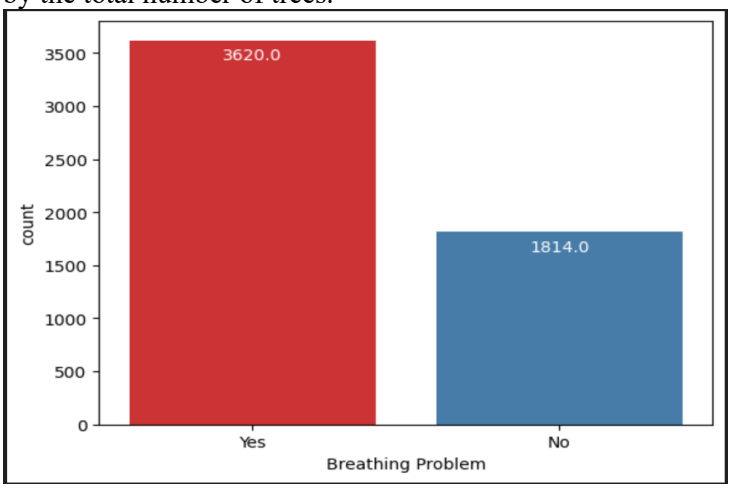


K Nearest Neighbors:

The oldest supervised machine learning algorithm for classification is KNN,which class if given instance According to the majority of categories among its k-nearest neighbours in the dataset.The distance Between the item to be categorized and every other item in the data set is calculated by the algorithm.

RandomForest: This classifer is a meta-estimator that adapts to decision trees on the dataset's different sub-samples and utilizes the average to increase the model'spredicted accuracy and controloverftting.In most

circumstances,this randomforest classifer seems to be more accurate than decision trees, and it also minimizes overftting.At the RandomForestlevel,average overall the trees is the final feature importance.The feature's importance sum value on each tree is numerically calculated and divided by the total number of trees.



RESULTAND DISCUSSION

To evaluate the effectiveness of the Machine Learning algorithms applied in this experiment,we decided to adopt the Accuracy,Mean squared error,Precision,Recall and F-Measure which are widely used in domains such as information retrieval,machine learning and other domains that involve binary classifcation.

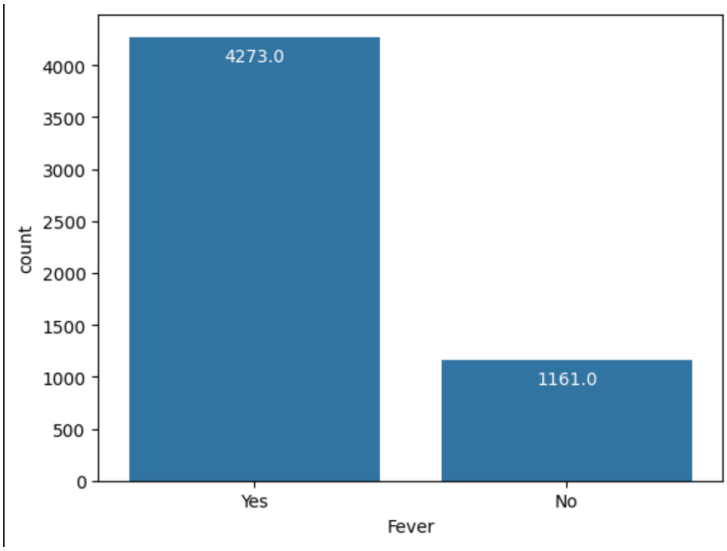


Figure shows that all of the algorithms performed well during the training process,owing to the fact that the Hyperparameters had been fine-tuned.However,there were minor discrepancies in accuracy,as indicated by the blue bar.The Random Forest Tree method has the best accuracy of all the algorithms,with a score of 98.39 percent.Additionally,R2 scores, meansquared errors and ROC scores are plotted in the bar chart using red, green and purple respectively.With 98.37 percent accuracy,the KNN is then extmost acceptable algorithm to use.

Model Performance and Results:

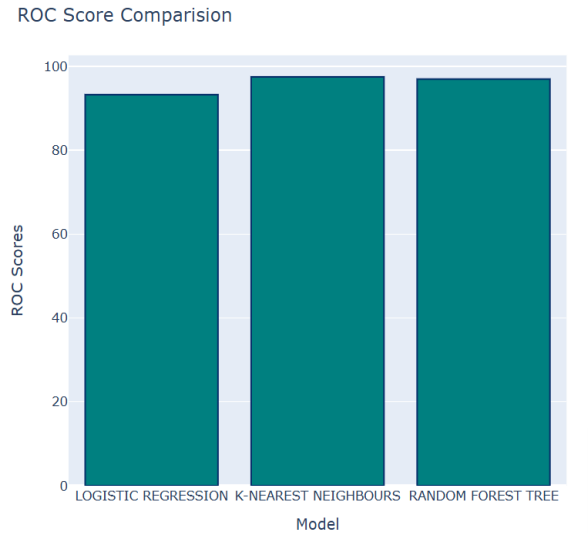
We have all the graphs for comparison but are not getting displayed .Please, check it in google collab.

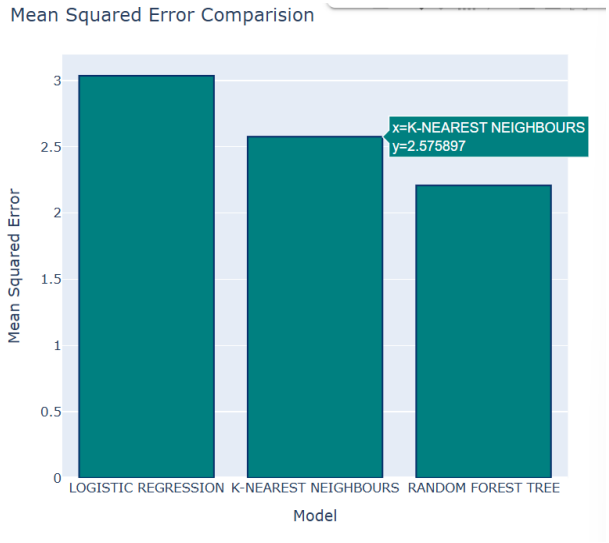
The project found significant differences in the performance of the various models. The key performance metrics are summarized below:

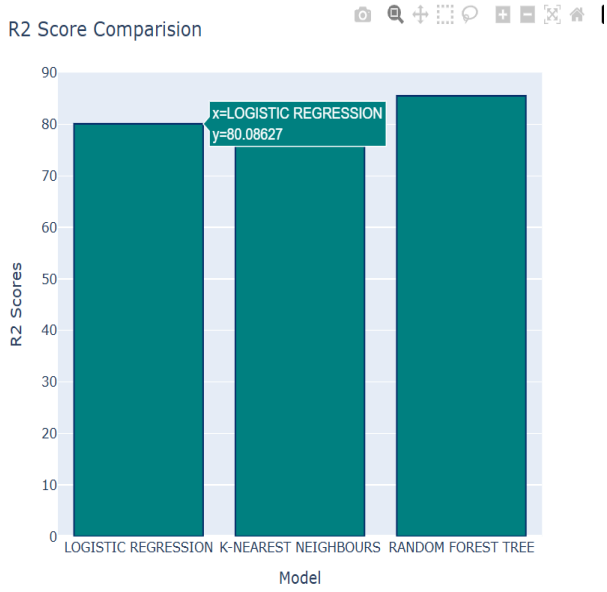
Model	Accuracy	ROC AUC (%)	MSE (%)	R-squared (%)
Logistic Regression	97.03%	93.23%	3.04%	80.09%
K-Nearest Neighbors (KNN)	98.37%	97.47%	2.58%	83.10%

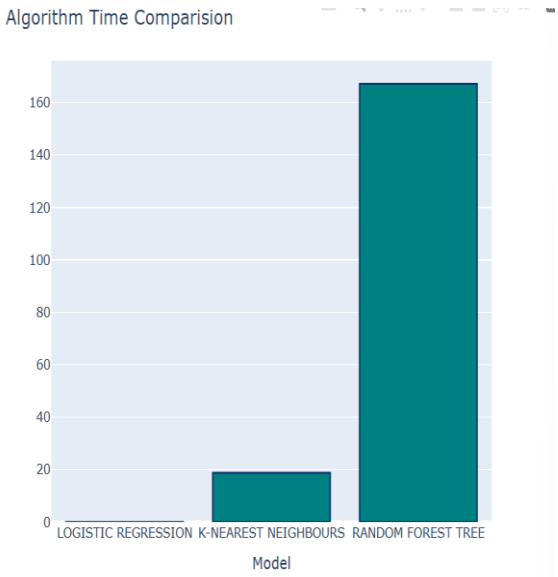
Model	Accuracy	ROC AUC (%)	MSE (%)	R-squared (%)
Random Forest	98.39%	96.94%	2.21%	85.52%
XGBoost	98.46%	97.15%	2.48%	83.71%
Decision Tree	98.46%	97.15%	2.48%	83.71%
Deep Neural Network (DNN)	98.36%	97.47%	2.58%	83.10%
Gaussian Naive Bayes	75.27%	84.54%	25.11%	-64.74%

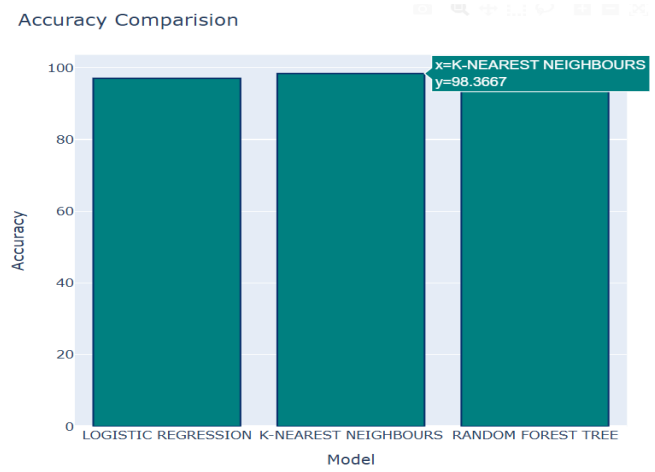
Export to Sheets











Based on these results, the **XGBoost** and **Decision Tree** models achieved the highest accuracy (98.46%). The **Random Forest** and **K-Nearest Neighbors** models also performed exceptionally well across multiple metrics. The **Gaussian Naive Bayes** model, however, had significantly lower performance.

In terms of computation time, the Logistic Regression model was the fastest (0.058 seconds), while the Random Forest model was the slowest (137.77 seconds).

Conclusion and Future Work:

The project successfully demonstrated that machine learning models can be highly effective in predicting COVID-19 diagnosis. The models developed as part of this project, particularly XGBoost, Decision Tree, and Random Forest, showed high accuracy and strong performance metrics.

The authors propose several areas for future work, including:

- **Feature Improvement:** Exploring dimensionality reduction or creating new interaction terms to enhance model performance.
- **Hyperparameter Refinement:** Using advanced optimization techniques like Grid Search or Bayesian optimization to find better hyperparameters.
- **Model Fusion:** Creating ensemble models by combining predictions from multiple models to improve robustness and precision.
- **Diverse Dataset Evaluation:** Testing the models on different datasets to assess their adaptability and prevent overfitting.
- **Exploring More Complex Models:** Investigating more advanced techniques like deep learning to further improve performance.

References:

[1]: de Moraes Batista, A. F., Miraglia, J. L., Rizzi Donato, T. H., & Porto Chiavegatto Filho, A. D. (2020). COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. https://doi.org/10.1101/2020.04.04.20052092

[2]: Z. Meng, M. Wang, H. Song, S. Guo, Y. Zhou, W. Li et al., ”Development and utilization of an intelligent application for aiding COVID-19 diagnosis”, medRxiv, 2020. https://www.medrxiv.org/content/10.1101/2020.03.18.20035816v1