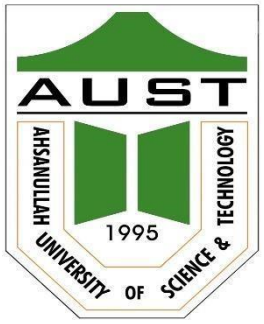

Ahsanullah University of Science & Technology
Department of Computer Science and Engineering

Course No: CSE3208
Course Title: Introduction to Artificial Intelligence Lab

PROJECT REPORT ON

COVID INFECTED PREDICTION

Submitted To:

Mr. Mohammad Marufur Rahman

Mr. Md. Aminur Rahman

Submitted By: A1(Group 2)

Zarin Tasnim	20200104007
Nusaiba Rafiq Surovi	20200104011
Azizah Mamun Abha	20200104021

Date of Submission: 20th August 2023

Experimental Domain:

The experimental domain for your COVID-19 infection probability prediction dataset encompasses a multidisciplinary approach. It involves collecting and preprocessing a diverse range of data, including demographic information, medical history, and social interactions. By leveraging machine learning algorithms such as decision trees, neural networks, and ensemble methods, you aim to create a robust predictive model. Through rigorous experimentation, you seek to optimize the model's accuracy, sensitivity, and specificity, contributing to the development of an effective tool for early identification and intervention of potential COVID-19 cases. This research not only advances the field of healthcare analytics but also holds significant societal implications by aiding in risk assessment and resource allocation during ongoing and future pandemics.

Description of Dataset:

This dataset contains information about individuals' attributes and medical conditions, with a focus on predicting the occurrence of a covid infected. The dataset comprises 5 features, each capturing different aspects of an individual's profile and health status. The target variable, 'Infected', 'Not-infected'," indicates whether an individual has experienced a Infected(1) or Not-infected (0).

Dataset Features:

1.Fever :Categorical feature indicating the body temperature in Fahrenheit .
2.Body Pain: Binary target variable indicating whether the individual has had body pain (1) or not (0).
3.Age: Numerical feature representing the age of the individual in years.
4.Runny Nose: Binary target variable indicating whether the individual has had a runny nose (1) or not (0).
5.Diff Breath: Binary target variable indicating whether the individual has had difficulties in breath (1) or not (0).
6.Infected :Binary target variable indicating whether the individual has had a Infected (1) or not Infected (0).

	fever	bodyPain	age	runnyNose	diffBreath	infected
count	798.000000	798.000000	798.000000	798.000000	798.000000	798.000000
mean	99.916106	0.468672	50.482456	0.505013	-0.018797	0.536341
std	1.170141	0.499331	28.789402	0.500288	0.816024	0.498990
min	98.002426	0.000000	1.000000	0.000000	-1.000000	0.000000
25%	98.912755	0.000000	25.000000	0.000000	-1.000000	0.000000
50%	99.835273	0.000000	50.000000	1.000000	0.000000	1.000000
75%	100.957035	1.000000	75.000000	1.000000	1.000000	1.000000
max	101.993067	1.000000	100.000000	1.000000	1.000000	1.000000

```
dataset.corr()
```

	fever	bodyPain	age	runnyNose	diffBreath	infected
fever	1.000000	0.034580	-0.009786	-0.026106	0.007453	0.002886
bodyPain	0.034580	1.000000	-0.068030	0.000629	-0.027621	-0.048300
age	-0.009786	-0.068030	1.000000	0.058939	0.039962	-0.007380
runnyNose	-0.026106	0.000629	0.058939	1.000000	-0.007452	0.024400
diffBreath	0.007453	-0.027621	0.039962	-0.007452	1.000000	-0.012187
infected	0.002886	-0.048300	-0.007380	0.024400	-0.012187	1.000000

ML Models:

Here, 6 machine learning models have been prepared using 6 different types of classifiers to evaluate the performances of the models. For the tuning of hyperparameters, a parameter grid has been implemented. Using this grid, a grid search has been performed for every type of classifier and the job of the grid search is to find the best values for the hyperparameters for each classifier.

Logistic Regression: It's particularly effective when the relationship between the features and the target variable is nonlinear. Key features of Logistic Regression include Binary and Multi-Class Classification, Linear Decision Boundary, Probability Interpretation.

- Hyper parameter Tuning: For optimization, the hyperparameter **solver** has been set to **liblinear** & **blogs** as the datasets we are working with are small in size compared to general large datasets. For the Inverse of regularization strength, **C** has been set to **0.01, 0.1, 1, 2** respectively.

Support Vector Classification: It is particularly good at solving binary classification problems, which require classifying the elements of a data set into two groups.

key features are Margin Maximization, Support Vectors' ability to handle nonlinear data.

- Hyperparameter Tuning: The regularization hyperparameter **C** has been set to **0.01,0.1** respectively and the type of kernels that have been used are **linear** and **rbf**.

Gaussian Naive Bayes: It's based on the principles of Bayes' theorem and assumes that features are normally distributed within each class. Key features of Gaussian Naive Bayes include Independence Assumption, Gaussian Distribution Training and Prediction, Suitable for continuous features.

- Hyperparameter Tuning: For Prior probabilities of the classes hyperparameter **priors** have been set to **None** and the value of **var_smoothing** which is the portion of the largest variance of all features that is added to variances for calculation stability has been set to **default**.

K Nearest Neighbors: In KNN, the prediction for a new data point is made by considering the labels or values of its nearest neighbors in the training dataset. The key features of this model are Distance Metric, Scalability, Voting (Classification), or Averaging (Regression).

- Hyperparameter Tuning: For the number of neighbors to use for k-neighbors queries, the hyperparameter **n neighbors** has been assigned values of 3,5,7,9 respectively.

Random Forest Classifier: Random Forest is an ensemble learning method that combines multiple individual decision trees to create a more robust and accurate model. Key features are Ensemble of Decision Trees, Random Subsampling.

- Hyperparameter Tuning: As for the number of trees in the forest, the hyperparameter **n_estimators** have been set to 100,200,300,400 respectively. Along with that, the maximum depth of the trees has been set to None, 5, and 10,15 respectively.

Decision Tree: A decision tree is a supervised machine learning algorithm that makes decisions or predictions represented in a tree-like structure. Decision trees are useful for both classification and regression tasks. Some key features are: Customizable Parameters, Handling Missing Values, Pruning, Visualization.

- Hyper parameter Tuning: The maximum depth of the decision tree has been set to None, 5, 10,15 respectively.

Performance Analysis:

A table of performance analysis using the best hyperparameters is provided below-

Model	Accuracy	F1 Score	Precision	Recall Score
Logistic Regression	**52%**	**66%**	**53%**	**88%**
SVC	**92%**	**90%**	**89%**	**90%**
Gaussian Naive Bayes	**79%**	**73%**	**70%**	**76%**
K Nearest Neighbors	**75%**	**77%**	**76%**	**79%**
Random Forest Classifier	**100%**	**100%**	**100%**	**100%**
Decision Tree	**100%**	**100%**	**100%**	**100%**

Description:

Having both Random Forest and Decision Tree models achieve 100% accuracy on your project is quite rare and might indicate overfitting. Both Random Forest and Decision Tree are popular machine learning algorithms. Decision Trees are simple to understand and can capture complex relationships in data. However, they are prone to overfitting when the tree becomes too deep.

Random Forest is an ensemble of Decision Trees that reduces overfitting by averaging the predictions of multiple trees. It combines their results to create a more robust and accurate model. Random Forest also handles features differently in each tree, reducing the likelihood of overfitting to specific features.

While high accuracy might seem promising, it's important to verify if your models are indeed performing well on unseen data, perhaps by using cross-validation or a separate test set. If your models are genuinely achieving 100% accuracy, it's worth considering if your dataset is too small, if there's data leakage, or if there's a mistake in the evaluation process.