Name - Vishal Adhav
BU ID - B01007320

Project 1

Methodology

1. **Data Preprocessing:**
    a. **Data Loading:**
    According to a dataset, there are survey replies for various states in the U.S. taken over three days. A particular day's survey result is captured in each row and 'tested_positive' which is the target column means the proportion of confirmed cases on day three.

    b. **Feature Selection and Normalization:**
    A dataset composed of diverse attributes like COVID-19 state-wide statistics, symptom observations, mental wellness indicators etc. was present in it.
    Since it had no significance towards predicting purposes, the id column was discarded.
    StandardScaler was used to normalize the features, so that differences in their orders of magnitude would not affect how the model is trained.

    c. **Model Structure (Deep Neural Network):**
    A fully connected DNN was used for the prediction task. The architecture is designed as follows:

    Input Layer: The input dimension equals the number of features.
    Hidden Layers:
    Layer 1: 128 neurons, ReLU activation, Dropout (0.2)
    Layer 2: 64 neurons, ReLU activation, Dropout (0.2)
    Layer 3: 32 neurons, ReLU activation, Dropout (0.2)
    Output Layer: 1 neuron for regression.
    Activation Function
    In order to achieve non-linearity, the ReLU activation function performed well in all hidden layers.
    Dropout
    Initially set at 0.3 but reduced to 0.2 to avoid underfitting.

    d. **Training Configuration:**
    Loss Function: This is the loss function that has been chosen to be mean squared error (MSE) which is appropriate for regression tasks
    Optimizer: Adam optimizer was used with a learning rate of 0.001.
    Epochs: The model was trained during 1000 epochs.

Batch Size: For training it has taken away from a batch size being used of 64 only.

e. **Hyperparameter Tuning:**
Several hyperparameters were experimented with, including:
Learning Rate: Tested 0.001 and 0.0001; settled on 0.001 due to faster convergence.
Batch Size: Experimented with 32, 64, and 128. The model performed best with a batch size of 64.
Dropout Rate: Initially set at 0.3 but reduced to 0.2 to avoid underfitting.

Empirical Results and Evaluation

1. **Training Process:**
The training process consisted of running the model for 1000 epochs. Below are some key observations during training:

Training Loss: The loss decreased steadily during the first 500 epochs and stabilized after that. The learning rate and dropout settings helped control overfitting, while the MSE loss provided a clear indicator of performance.

Validation Loss: The model achieved a validation loss of 1.1212, which shows reasonable accuracy in predicting the percentage of positive cases on the third day.

Epoch	Training Loss	Validation Loss
100	**0.0045**	**0.0051**
200	**0.0032**	**0.0039**
500	**0.0014**	**0.0021**
1000	**0.0008**	**0.0017**

2. **Model Evaluation:**
The assessment of the model was conducted utilizing the Mean Squared Error (MSE) on the validation set. The absolute validation MSE which amounted to 1.1212 implies that the model was able to give an adequate level of precision while predicting for the fraction of positive cases. Nonetheless, there could be a chance for betterment according to the training loss because it kept on reducing during even the 1000th epoch.

Name - Vishal Adhav
BU ID - B01007320

3. **Predictions:**
   Predictions were made on the test data, and the results were saved to a CSV file. The model demonstrated good predictive ability on unseen data, which was confirmed through low validation loss.

- Conclusion:

Developed in the project is a program for predicting the percentage of positive COVID-19 cases using past survey data with the aid of Deep Neural Network (DNN). By being aware of how the data were preprocessed, how the model was designed and which hyperparameters were tuned, we managed to achieve a validation loss rate of 1.1212.