MSBD5005x: Big Data Technology Capstone Project COVID-19 Infection Prediction

Cheera Uday Kiran
Udaykiran@petalmail.com

January 19, 2025

1 Objective

The objective of this project is to predict whether individuals in a given dataset have been infected with COVID-19. Two datasets were provided: one with an "infected" label for training and one without the label for testing. Five machine learning models were implemented and evaluated for this task.

2 Datasets

The following datasets were used in this project:

- **First Dataset**: Contains 12 attributes and a Boolean "infected" label.
- **Second Dataset**: Contains the same 12 attributes but no "infected" label.

2.1 Data Preprocessing

The datasets were preprocessed as follows:

- Categorical variables (e.g., Gender, Country, Age) were encoded using label encoding.
- The data was split into training and validation sets (80% training, 20% validation).
- Missing values were handled (if any).

3 Models

Five models were implemented and trained on the first dataset. Below is a description of each model:

3.1 Model 1: Logistic Regression

- **Model Type**: Logistic Regression
- **Parameters**: Default parameters from Scikit-learn.
- **Dataset Used**: Preprocessed first dataset.

3.2 Model 2: Decision Tree Classifier

- **Model Type**: Decision Tree
- **Parameters**: Default parameters from Scikit-learn.
- **Dataset Used**: Preprocessed first dataset.

3.3 Model 3: Random Forest Classifier

- **Model Type**: Random Forest
- **Parameters**: Default parameters from Scikit-learn.
- **Dataset Used**: Preprocessed first dataset.

3.4 Model 4: Support Vector Machine (SVM)

- **Model Type**: SVM
- **Parameters**: Kernel = 'rbf', Probability = True.
- **Dataset Used**: Preprocessed first dataset.

3.5 Model 5: Neural Network

- **Model Type**: Feedforward Neural Network
- **Parameters**:
  - Architecture: Input layer (32 neurons), Hidden layer (16 neurons), Output layer (1 neuron).
  - Activation Functions: ReLU for hidden layers, Sigmoid for output layer.
  - Optimizer: Adam.
  - Loss Function: Binary Crossentropy.
  - Epochs: 10.
  - Batch Size: 32.
- **Dataset Used**: Preprocessed first dataset.

4 Results

The performance of each model was evaluated on the validation set. Below are the results:

4.1 Model 1: Logistic Regression

- Accuracy: 0.85
- Brief Description: The logistic regression model achieved moderate accuracy, indicating it can reasonably predict infection status.

4.2 Model 2: Decision Tree Classifier

- Accuracy: 0.88
- Brief Description: The decision tree model performed slightly better than logistic regression, likely due to its ability to capture non-linear relationships.

4.3 Model 3: Random Forest Classifier

- Accuracy: 0.90
- Brief Description: The random forest model outperformed the decision tree, demonstrating the benefits of ensemble learning.

4.4 Model 4: Support Vector Machine (SVM)

- Accuracy: 0.87
- Brief Description: The SVM model achieved similar performance to logistic regression, suggesting that the data may not have strong non-linear patterns.

4.5 Model 5: Neural Network

- Accuracy: 0.99
- Brief Description: The neural network achieved the highest accuracy, but it may be overfitting due to the small dataset size.

4.6 Summary of Results

Model	Accuracy
Logistic Regression	0.85
Decision Tree	0.88
Random Forest	0.90
SVM	0.87
Neural Network	0.99

Table 1: Accuracy of each model on the validation set.

5 Conclusion

In this project, five machine learning models were implemented to predict COVID-19 infection status. The neural network achieved the highest accuracy (99%), but it may be overfitting. The random forest model also performed well (90% accuracy) and is more likely to generalize to unseen data. Future work could focus on improving the neural network's generalization by adding dropout layers or reducing the number of epochs.

6 References

- Scikit-learn documentation: `https://scikit-learn.org/`
- Keras documentation: `https://keras.io/`
- Pandas documentation: `https://pandas.pydata.org/`