AML Project Report

NAME:	M. Awais Tariq
CLASS:	MS-AI
REG.ID:	23I-7606
SUBJECT:	AML
SUBMITTED TO:	Dr. Uzair Iqbal

Predictive Modeling for COVID-19 Patient Classification: Leveraging Automated Feature Engineering and Hyperparameter Optimization

Domain Background:

The COVID-19 domain highlights the critical importance of efficient resource allocation and patient management. It helps health authorities anticipate risk allocation in patients with COVID-19, prioritize resources, and use appropriate resources.

Objective: This paper investigates in detail the development of machine learning models with the intention of predicting COVID-19 risk classification based on patient characteristics, conditions and medical history.

Dataset Description:

The dataset was provided by the Mexican government. This dataset contains an enormous number of anonymized patient-related information including pre-conditions. The raw dataset consists of 21 unique features and 1,048,576 unique patients. The target variable, 'classification,' indicates COVID-19 test findings. Values 1-3 mean that the patient was diagnosed with covid in different degrees, 4 or higher means that the patient is not a carrier of covid or that the test is inconclusive.

ML Model Architecture:

1. **Data Preprocessing:** Load and preprocess the dataset for filling missing values, converting categorical variables into suitable format for modelling.

2. **Feature Engineering:** Complex relationships and patterns are captured by generating new features from the data set using FeatureTools.

3. **Feature Selection:** The feature extractor prepares a matrix of suitable datasets after which applies diverse strategies of decreasing dimensionality such as correlation analysis, variance thresholding, or recursive feature removal.

4. **Ensemble Classification algorithm :** A large wide variety of different type algorithms are used to study from conditionally unbiased subsets created by way of the characteristic extractor . These are Gradient Boosting, Random Forests or XGBoost amongst others.

5. **Hyperparameter Optimization:** Optuna is used with a purpose to optimize hyper-parameters for models at the same time as ensuring robustness in overall performance.

6. **Evaluation:** Model performance is evaluated using accuracy metrics including precision, classification report and F1-score for comprehensive evaluation of predictive capability.

7. **Final Model Selection:** Based on assessment of metrics in the classification report decided on best models, are deployed onto unseen statistics units for prediction functions.

Contributions:

Feature Engineering:

1. **EntitySet Creation:** Employed the library Featuretools to create an EntitySet 'covid19' and then added the root entity, the main dataframe 'data' with COVID-19 patients data.

2. **DFS:** Automatic generation of numerous new features was carried out employing Deep Feature Synthesis and transformation primitives like 'add_numeric', 'multiply_numeric'. Thus, the dataset was enriched with a considerable number of engineered features that were derived from the existing features.

Feature Selection Algorithm:

1. Firstly, used the correlation matrix to check & remove correlated features, which were greater than 0.85, so as to improve model learning.

2. Then removed features, which had variance value less than 0.1.

3. Thirdly, used the feature importance function of the various classifiers (GBM, XGB, Rand Forest, etc.) to get important features, and fitted an ensemble classifier on them..

4. These important features were then recursively selected using RFE based on their importance in model tuning.

5. finally, normalized all the features in between 0 and 1, selected the features which were above a certain threshold.

Hyperparameter optimization:

In this project we used Optuna library for hyperparameter optimization of GBM & AdaBoost classifiers.

Refrences:

1. COVID-19 Dataset (kaggle.com)

2. COVID - 19: KFold and Random Forest Classifier (kaggle.com)

3. Comparing Tree Based Models with Covid 19 Dataset (kaggle.com)

4. What is Featuretools? — Featuretools 1.30.0 documentation (alteryx.com)