**Project Report – Part 2**

**CLASSIFICATION ALGORITHM**:

**Random Forest Classifier**

Random Forest model is the chosen choice for classifying disease for the given dataset. It is an ensemble tree-based machine learning algorithm which uses a set of decision trees trained on a randomly selected sample with replacement from training data.

The classification results from all independent trees in the Random Forest model are considered to reach final prediction for each data point in test set through voting mechanism.

**PARAMETER SETTINGS:**

All parameters not explicitly mentioned below are set to their default value as per their documentation page.

Final parameter setting after hyper parameters tuning using GridSearchCV approach are:

{'criterion': 'gini',

'max\_depth': None,

'max\_features': 'auto',

'min\_samples\_leaf': 1,

'min\_samples\_split': 2,

'n\_estimators': 50}

**CONFUSION MATRIX:**

Disease and its associated numerical label:

|  |  |
| --- | --- |
| **Disease** | **Label** |
| ﻿Healthy | 0 |
| 21OHD-NC | 1 |
| 21OHD-C Disease | 2 |
| 3beta-HSD like | 3 |
| 3beta-HSD | 4 |
| 11beta-OHD | 5 |
| 17OHD | 6 |
| Aldosteron sentaz | 7 |
| POR | 8 |
| NonCAH PAI | 9 |

Confusion Matrix:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| **0** | 140 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| **1** | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | 0 | 0 | 18 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **3** | 1 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| **4** | 0 | 0 | 1 | 1 | 5 | 0 | 0 | 0 | 0 | 0 |
| **5** | 0 | 0 | 0 | 0 | 0 | 18 | 0 | 0 | 0 | 0 |
| **6** | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 |
| **7** | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| **8** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **9** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |

**ACCURACY:**

Accuracy Percentage = 196/207 = 94.686%