**Experimental Analysis**

Dataset details and the result analysis is represented in this section.

**Dataset Details**

This dataset contains the information of 606 persons. It includes data about peoples including risk factors of developing stroke that may cause stroke. This dataset has been created from a direct questionnaire to people who have recently developed stroke, or who are still not developed the stroke but having few or more risk factors of stroke. The data has been collected from the patients using direct questionnaire from diﬀerent hospital in Sylhet, Bangladesh. We have collected the information from Sylhet Woman Medical College and Hospital, Jalalabad Ragib Rabeya Medical College and hospital. The description of dataset is given below.

|  |  |  |
| --- | --- | --- |
|  | Number of Attributes | Number of Instances |
| Risk Factors Dataset | 15 | 606 |

# Table 1:Description of Dataset

|  |  |
| --- | --- |
| Attributes | Values |
| Age | 1.25-34, 2.35-44, 3.45-54,4.55-65,5.65< |
| Gender | 1. Male 2. Female |
| Systolic BP | 1.120>, 2.120-139, 3. 140-160, 4.160< |
| Diastolic BP | 1.180>, 2.80-95, 3.95< |
| Diabetes | 1. No, 2. Yes |
| Ischemic Heart Disease | 1. No, 2. Yes |
| Family History of stroke | 1. No, 2. Yes |
| Alcoholism | 1. No, 2. Yes |
| Less Physically Active | 1. No, 2. Yes |
| Smoking | 1. No, 2. Yes |
| Stress and depression | 1. No, 2. Yes |
| Saturated Fat↑ () | 1. No, 2. Yes |
| Fibre↓ () | 1. No, 2. Yes |
| Chronic Kidney Disease (CKD) | 1. No, 2. Yes |
| Class Attribute | 1. Stroke, 2. Non-stroke |

# Table 2:Description of Attribute

The data pre-processing has been conducted by handling the missing values following the technique of ignoring the tuples with incomplete values. After pre-processing, 606 instances have been remained in total. Among them, 451 are positive values and 155 are negative values. The detail description of the attributes is shown in Table 2. Two class variables are used to ﬁnd whether the patient is having a risk of developing of stroke (positive) or not (negative).

Figure 1:Class Attributes Distribution

**Result Analysis**

Performance of diﬀerent Data Mining techniques on our dataset with detailed accuracy information is represented in the following tables. Although Nave Bayes classiﬁer is one of the most popular algorithms for data prediction, in case of our dataset, the accuracy of it was the lowest for both the Cross-validation method and also for the Percentage split. However, the best result was achieved by using RandomForest decision tree where using 10-fold cross validation 84.16% instances were classiﬁed correctly and using percentage split technique it could classify 80.99% of the instances correctly. In Table 3 to Table 18 we have depicted the detail analysis result. We have found the correctly classiﬁed instances and incorrectly classiﬁed instances for each algorithm.

|  |  |  |
| --- | --- | --- |
|  | Number of Instances | Percentage |
| Correctly classiﬁed Instances | 510 | 84.16% |
| Incorrectly Classiﬁed Instances | 96 | 15.84% |

# Table 3: Performance Results from RandomForest decision tree using (Cross Validation)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-measure |
|  | 0.665 | 0.098 | 0.701 | 0.665 | 0.682 |
|  | 0.902 | 0.335 | 0.887 | 0.902 | 0.895 |
| Weighted Average | 0.842 | 0.275 | 0.839 | 0.842 | 0.840 |

# Table 4: Detailed Accuracy by class from RandomForest decision tree using 10-fold Cross Validation technique

|  |  |  |
| --- | --- | --- |
|  | Number of Instances | Percentage |
| Correctly classiﬁed Instances | 98 | 80.99% |
| Incorrectly Classiﬁed Instances | 23 | 19.0083% |

# Table 5: Performance Results from RandomForest decision tree using Percentage Split

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-measure |
|  | 0.585 | 0.075 | 0.800 | 0.585 | 0.676 |
|  | 0.925 | 0.415 | 0.813 | 0.925 | 0.865 |
| Weighted Average | 0.810 | 0.300 | 0.809 | 0.810 | 0.801 |

# Table 6: Detailed Accuracy by class from RandomForest decision tree using - Percentage Split (80:20)

|  |  |  |
| --- | --- | --- |
|  | Number of Instances | Percentage |
| Correctly classiﬁed Instances | 449 | 74.09 % |
| Incorrectly Classiﬁed Instances | 157 | 25.91% |

# Table 7: Performance Results from Naïve Bayes Algorithm using (Cross Validation)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-measure |
|  | 0.303 | 0.109 | 0.490 | 0.303 | 0.375 |
|  | 0.891 | 0.697 | 0.788 | 0.891 | 0.837 |
| Weighted Average | 0.741 | 0.546 | 0.712 | 0.741 | 0.718 |

# Table 8: Detailed Accuracy from by class Naïve Bayes with 10-fold Cross Validation technique

|  |  |  |
| --- | --- | --- |
|  | Number of Instances | Percentage |
| Correctly classiﬁed Instances | 82 | 67.77% |
| Incorrectly Classiﬁed Instances | 39 | 32.23% |

# Table 9: Performance Results from Naïve Bayes - Percentage Split

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-measure |
|  | 0.195 | 0.075 | 0.571 | 0.195 | 0.291 |
|  | 0.925 | 0.805 | 0.692 | 0.925 | 0.791 |
| Weighted Average | 0.678 | 0.558 | 0.651 | 0.678 | 0.622 |

# Table 10: Detailed by class Accuracy from Naïve Bayes– Percentage Split (80:20)

|  |  |  |
| --- | --- | --- |
|  | Number of Instances | Percentage |
| Correctly classiﬁed Instances | 489 | 80.69% |
| Incorrectly Classiﬁed Instances | 117 | 19.31% |

# Table 11: Performance Results from RandomTree decision tree Algorithm using (Cross Validation)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-measure |
|  | 0.677 | 0.149 | 0.610 | 0.677 | 0.642 |
|  | 0.851 | 0.323 | 0.885 | 0.851 | 0.868 |
| Weighted Average | 0.807 | 0.278 | 0.815 | 0.807 | 0.810 |

# Table 12: Detailed Accuracy from RandomTree using 10-fold Cross Validation technique

|  |  |  |
| --- | --- | --- |
|  | Number of Instances | Percentage |
| Correctly classiﬁed Instances | 95 | 78.51% |
| Incorrectly Classiﬁed Instances | 26 | 21.49% |

# Table 13: Performance Results from RandomTree using - Percentage Split

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-measure |
|  | 0.659 | 0.150 | 0.692 | 0.659 | 0.675 |
|  | 0.850 | 0.341 | 0.829 | 0.850 | 0.840 |
| Weighted Average | 0.785 | 0.277 | 0.783 | 0.785 | 0.784 |

# Table 14: Detailed Accuracy from RandomTree using - Percentage Split (80:20)

|  |  |  |
| --- | --- | --- |
|  | Number of Instances | Percentage |
| Correctly classiﬁed Instances | 451 | 74.42% |
| Incorrectly Classiﬁed Instances | 155 | 25.56% |

# Table 15: Performance Results from Support Vector Machine Algorithm using (Cross Validation)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-measure |
|  | 0 | 0 | ? | 0 | ? |
|  | 1 | 1 | 0.786 | 1 | 0.853 |
| Weighted Average | 0.744 | 0.744 | ? | 0.744 | ? |

# Table 16: Detailed Accuracy from Support Vector Machine using - Cross Validation

|  |  |  |
| --- | --- | --- |
|  | Number of Instances | Percentage |
| Correctly classiﬁed Instances | 80 | 66.12 % |
| Incorrectly Classiﬁed Instances | 41 | 33.88% |

# Table 17: Performance Results from Support Vector Machine Algorithm – Percentage Split

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-measure |
|  | 0 | 0 | ? | 0 | ? |
|  | 1 | 1 | 0.661 | 1 | 0.796 |
| Weighted Average | 0.661 | 0.661 | ? | 0.661 | ? |

# Table 18: Detailed Accuracy from Support Vector Machine using 10 Percentage Split (80:20)

For the more semantic view of the performance of used algorithms using both evaluation techniques are depicted in graphs. In Fig. 2, the performance of the algorithms using Cross-validation evaluation is depicted and in Fig. 3, the results from percentage split have been shown to represent the comparative accuracy of the used algorithms.

# Figure 2:Performance of Classiﬁcation Algorithms Using Cross-Validation Technique

# Figure 3:Performance of Classiﬁcation Algorithms Using Percentage Split Technique

From table 3 to 18 we can see that the decision tree gives the best accuracy performance according to our datasets. After seeing this we have analyzed our datasets by implementing the code of the decision tree in python. We implemented the code in python in two ways. First, we implemented the code with k-fold cross validation. For this we split the dataset into k-fold and calculate the percentage of the performance accuracy. Then we calculate the Gini Index of the attributes and attributes values for the dataset and create a terminal node value. Then we generate a decision tree by following the CART (Classification and Regression Tree) algorithm and find 80.33% the mean accuracy of the algorithm.