1. **Project 1: Diabetes Prediction from Medical Records**

**The following findings and results are based on the machine learning analysis of the provided dataset i.e. Diabetes EHR data. The applied machine learning approach demonstrates the application of 5 different machine learning models based on classification methods which are *Random Forest, Naïve Bayes, Logistic Regression, Support Vector Machine, K-Nearest Neighbor.***

1. **Data Exploration and Visualization**

**The provided EHR diabetic data from the National Institute of Diabetes and Digestive and Kidney Diseases consists information of female patients (aged ≥ 21) of Pima Indian Heritage. Upon initial inspection it was found that there are 550 entries (rows) and 10 variables (columns) in the dataset. The predictor variables are Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and Outcome, where Outcome has two classes 0 and 1, 0 for healthy and 1 for diabetic.**

**There are no missing data in this dataset as verified by following steps.**

**Some quick stats of these variables including mean, standard deviation, min and max is depicted below as a part of exploratory data analysis.**

**Some Visualization of these variables are also depicted below.**

***Fig 1: Distribution of Variables***

**The figures illustrated above provides the overview of variables distribution. Variables like *Glucose, BloodPressure, SkinThickness,* and *Outcome* have a bell-shaped formation indicating normal distribution. Rest of others follow slight left or right skewed but that is something to worry since it is expected with the medical nature of dataset**.

***Fig 2: Plot of Class Imbalance***

**The Class imbalance plot above illustrates the count of 0 (healthy) and 1 (diabetic) cases in the dataset. It can be seen that there are 350 cases of 0 but only 200 cases of 1 which represents a slight imbalance but we can address this during the evaluation process.**

***Fig 3: Box Plot of Variables***

**The box plot of variables above provides the snapshot of presence of any outliers within the dataset but so far, our data looks normal there is no extreme cases of outlier which can alter our findings.**

***Fig 4: Correlation Heatmap of the Variables***

**The Correlation heatmap above suggest that Glucose (0.5) have the highest or strongest positive correlation with the target variable outcome meaning higher glucose level are moderately associated with having diabetes. BMI and Age also suggest positive but small correlation. Since we are using a comprehensive approach for this analysis, we will be using all relevant variables as these medical biomarkers are indeed relevant for our study.**

1. **5 Machine Learning Models**

**To carry out with analysis, 5 machine learning classification methods were used; they are.**

1. **Support Vector Machine (SVM)**
2. **K-Nearest Neighbor (KNN)**
3. **Random Forest**
4. **Naïve Bayes**
5. **Logistic Regression**
6. **Averaged Results using 10-fold Cross Validation**

**A 10-fold Stratified Cross Validation was used in order to address the slight imbalance that was encountered. The results of the 10-fold cross validation are illustrated in the table below.**

1. **Tables and Figures of Classification F1-Scores and Accuracies**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **F1-Score** | **Precision** | **Recall** |
| **Naive Bayes** | **0.7636** | **0.6455** | **0.6968** | **0.6108** |
| **Random Forest** | **0.7618** | **0.6417** | **0.7005** | **0.6100** |
| **Logistic Regression** | **0.7600** | **0.6204** | **0.7098** | **0.5592** |
| **Support Vector Machine (SVM)** | **0.7509** | **0.6057** | **0.6876** | **0.5487** |
| **K-Nearest Neighbors (KNN)** | **0.7327** | **0.5952** | **0.6412** | **0.5629** |

**Table 1: Performance Comparison of Models**

**From the table above, we have accuracy, F1-score, precision and recall for all the models. These scores are further demonstrated graphically below to provide a visual overview.**

***Fig 5: Bar Plot of Accuracy Scores of all Models***  ***Fig 6: Bar Plot of F1-Scores of all Models***

1. **Conclusion and Discussion on the best Classification method and Classification Report**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | **F1-Score**  **(Weighted Avg)** | **Notes** |
| **Random Forest** | **1.0000** | **1.0000** | **Perfect scores but highly suspicious of overfitting!** |
| **KNN** | **0.7891** | **0.7835** | **Good performance** |
| **Logistic Regression** | **0.7691** | **0.7627** | **Decent, balanced** |
| **Naive Bayes** | **0.7727** | **0.7692** | **Decent, balanced** |
| **SM** | **0.7236** | **0.6879** | **Lowest performance** |

***Table 2: Classification Reports***

**Considering all the results from accuracy, F1 scores, classification report, etc. the best model for our data is the Naïve Bayes. It has an accuracy of 0.7636 and a F1 score with 0.6455 which in comparison is a better number given other models and has a balanced fit.**