Stroke Prediction Dataset

Creation of Decision Tree and Random Forest machine learning classification models,
Alessandro Sciorilli

The following report presents the methodology and results of applying Decision Tree and Random Forest machine learning classification models to a Stroke Prediction Dataset obtained from Kaggle. These models are designed to predict the probability of a patient to experience a stroke (labeled as 'y'). The prediction is based on a range of features, including gender, age, hypertension, heart diseases, marital status (ever_married), work type, residence type, average glucose level, BMI, and smoking status. Using these machine learning classification models can help healthcare professionals identify the most common factors associated with instances of stroke, enabling them to provide focused advice and medical treatments to patients.

Upon importing the required libraries and reading the CSV file into a Pandas dataframe, I conduct data exploration and remove all rows containing missing values. Specifically, to gain clearer insights into the features associated with stroke episodes, I decide to exclude all entries indicating **'unknown'** under the **'smoking status'** feature.

After analyzing the data through visualization, it becomes immediately clear that the dataset is skewed towards individuals who haven't experienced a stroke. By using a **stripplot**, it is possible to visualize how instances of stroke are more widely spread across higher age groups. To accurately evaluate the distribution of stroke occurrences across different categorical variables, I compute the percentages. This approach normalizes the impact of stroke incidents within each category, making it easier to compare them meaningfully despite variations in category sizes.

The insights gained from this visualization (depicted in **Plot 4**) reveal that strokes are slightly more common among males **(2.1%)** than females **(1.8%).** Additionally, strokes occur more frequently among self-employed individuals **(3.29%)** compared to other employment categories (**1.57%** for government jobs and **1.64%** for positions in private companies). Urban residents **(1.91%)** also experience slightly more strokes than their rural counterparts **(1.86%).** Interestingly, former smokers **(2.54%)** tend to experience strokes more often than current smokers **(1.80%)**.

I perform one-hot encoding on the categorical variables**: 'gender,' 'ever_married,' 'work_type,' 'Residence_type,' and 'smoking_status.'** The correlation matrix reveals a slight positive correlation between stroke and the variables: age **(+15%)**, hypertension **(+8%)**, heart_disease **(+11%)**, and average glucose level **(+8%)**. Plotting the distribution of strokes, I observe again that the dataset is strongly unbalanced toward the individuals with no stroke. To avoid biased model performance, I rebalance my dataset by randomly selecting **500** samples from each class (**0 = 'No Stroke'** and **1 = 'Stroke'**). Moreover, I set a random seed equal to 15 in order to keep the sequence of random numbers consistent.

As the next step, I separate the features (X) from the label **'stroke'** (y) and split my dataset into training and test sets using a **70/30** ratio. Afterward, I rescale my data through standardization. I proceed creating and training the **Decision Tree** classifier model:

```
dt = DecisionTreeClassifier(random_state=15)
          dt.fit(X_train, y_train)
```

Subsequently, I perform **5-Fold Cross Validation**, obtaining an average accuracy of 0.65.

Exploring the different parameters of the model, it is specified that the criterion is **'Gini'**. Gini is a measure used by the cart algorithm to measure impurity, calculated as:

```
Gini impurity = 1 - (probability of 'yes')² - (probability of 'no')²
```

From the classification report, for the **'no stroke'** class, the **precision** is **0.69**, which means that out of all instances predicted as **'no stroke'**, **69%** are correct. For the **'stroke'** class, the precision is **0.78**, indicating that **78%** of instances predicted as **'stroke'** are correct. **Accuracy** is **74%**, **F1-score** is **0.73** for the **'no stroke'** class and **0.74** for the **'stroke'** class. Overall, the metrics indicate that the model is making reasonably accurate predictions, although further analysis might be needed to understand factors influencing the performance and possible ways to improve it.

I explore which feature of the dataset is more likely to indicate the occurrence of a stroke. Using the function **'feature_importances'**, I rank the features in order of impurity inside a dataframe. As a result, the feature **'age'** has the worst impurity score (highest value). Therefore, 'age' will be the root of the tree, **'avg_glucose_level'** and **'bmi'** will be the two primary sub-children, and so on.

Finally, I create and train the **Random Forest** classifier.

```
rf = RandomForestClassifier(random_state=15)
      rf.fit(X_train, y_train.ravel())
```

Using **5-Fold Cross Validation**, I achieve an average accuracy of **0.73**. The **classification report** for the random forest demonstrates more robust metrics values. Specifically, the **precision** is **0.74** for the **'no stroke'** class and **0.80** for the **'stroke'** class. The **accuracy** is **0.77**, and the **f1-score** is **0.76** for the **'no stroke'** class, whereas it reaches **0.79** for the **'stroke'** class. The features importance for the Random Forest confirms the same ranking as before, highlighting **'age'**, **'avg_glucose_level,'** and **'bmi'** as the most impactful features.

In conclusion, employing Decision Tree and Machine Learning algorithms to predict stroke probability, I identified age, blood glucose levels, and BMI as the most influential features. This information could empower doctors to offer tailored recommendations to patients, suggesting more frequent health assessments as age advances. Additionally, closely monitoring glucose levels and maintaining a healthy BMI becomes crucial for effective preventive measures against the occurrence of strokes.

Dataset link: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset