Diabetes Prediction

Github link : https://github.com/R4THUR/Final-Assignement-DA

ALGORITHM DESCRIPTION :

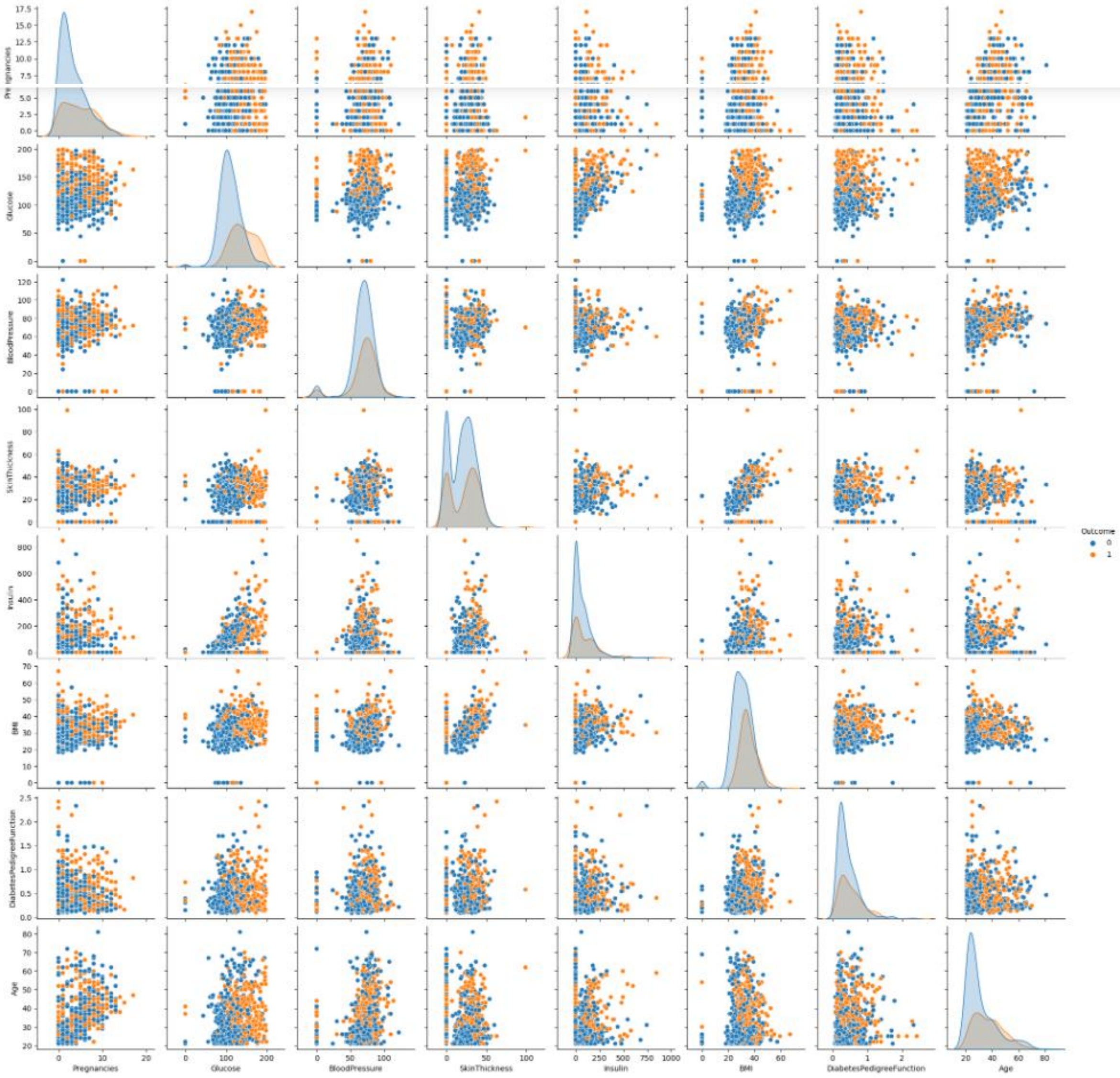
The algorithm utilized for this solution is K-Nearest Neighbors (KNN). KNN is renowned for its simplicity and effectiveness in classification tasks. It operates by classifying data points based on the majority class among their nearest neighbors in feature space. Unlike some algorithms, KNN doesn't involve explicit training; rather, it memorizes the training data and makes predictions based on the similarity of new data points to the existing ones. This makes it a non-parametric algorithm, meaning it doesn't assume any underlying probability distributions of the data. The key hyperparameters of KNN include 'k' (the number of neighbors to consider) and the weight function used in prediction, which can be either 'uniform' or 'distance'-based.

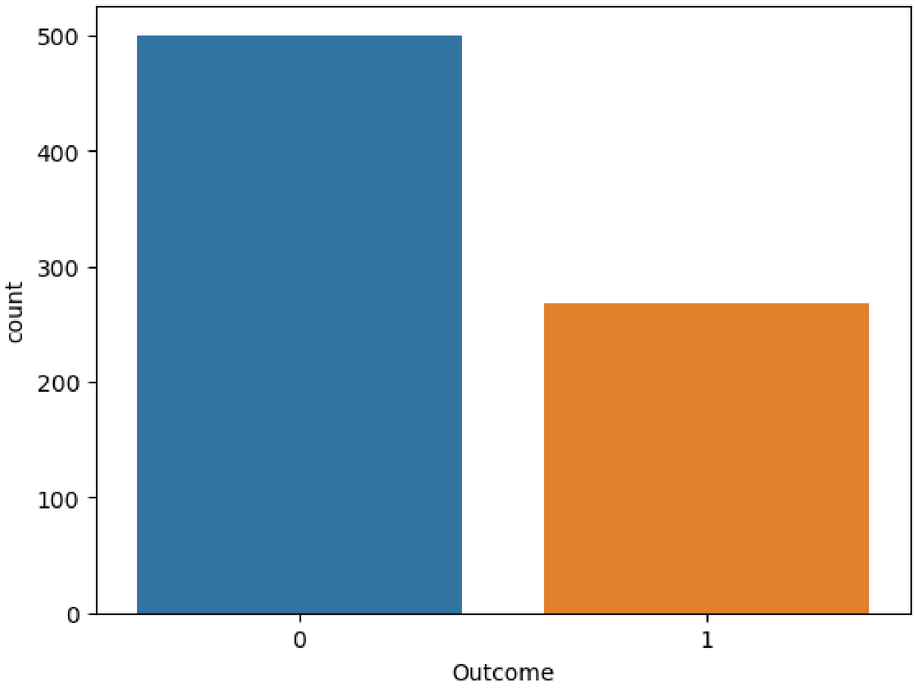
MODEL PERFORMANCE :

To enhance model performance, hyperparameter tuning was conducted using Grid Search with Cross-Validation. This method exhaustively searches through a specified hyperparameter grid to find the combination that yields the best performance. Cross-validation is employed to evaluate each hyperparameter combination, providing a robust estimate of model performance. In this case, the hyperparameters tuned were 'n_neighbors', ranging from 1 to 30, and 'weights', offering two options: 'uniform' and 'distance'. Here we got the results : Best hyperparameters: {'n_neighbors': 15, 'weights': 'distance'}. By systematically exploring these hyperparameters, the model can adapt better to the underlying patterns in the data, thereby improving its predictive accuracy.

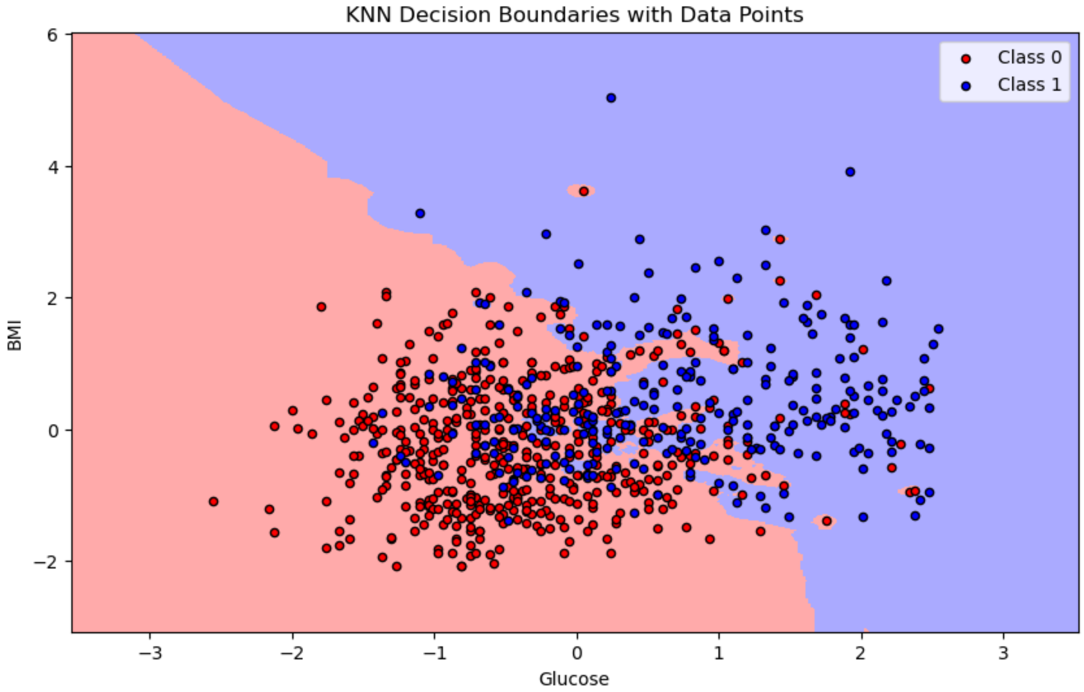
GRAPHS :

Graphical representations are instrumental in comprehending the model's training process and evaluating its performance. The visualization begins with an exploratory data analysis (EDA) phase, including plots such as pair plots and count plots. These visualizations provide insights into the distribution of features and the target variable ('Outcome').



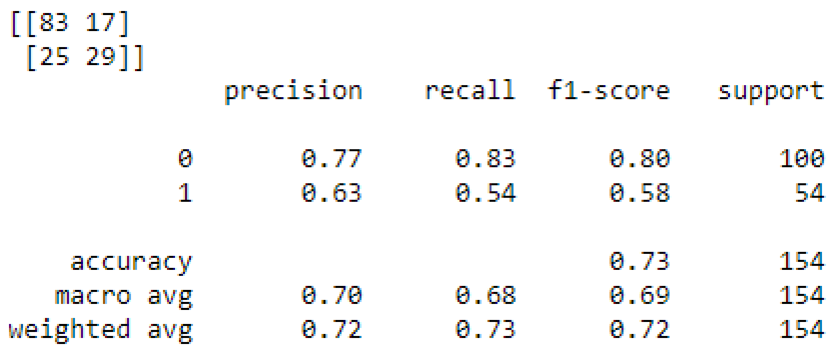


Subsequently, the KNN decision boundaries are visualized along with the data points to understand how the model categorizes different feature combinations. In this visualization, the KNN decision boundaries divide the feature space into regions predicted as diabetic (blue) or non-diabetic (red). Data points from the dataset are plotted based on their feature values (Glucose and BMI) and actual class labels, distinguished by color. Lighter shades of blue indicate regions predicted as diabetic, while lighter shades of red indicate non-diabetic regions. The legend clarifies the color-class correspondence. This visualization aids in understanding how the KNN classifier categorizes different regions of the feature space, providing insights into its predictive behavior.



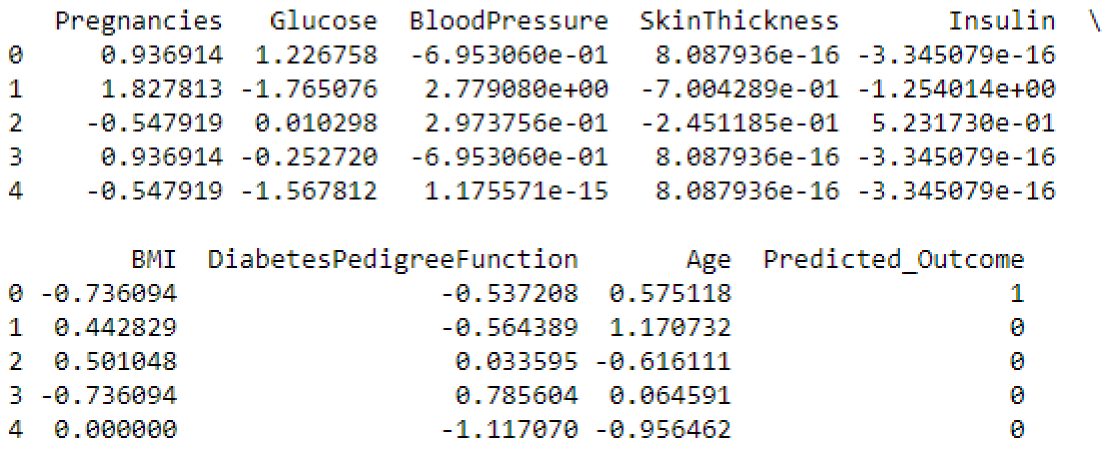
RESULTS :

The output provided includes a confusion matrix and a classification report, followed by individual predictions. The confusion matrix reveals the model's performance in classifying instances into true positives, false positives, true negatives, and false negatives. Specifically, out of 100 non-diabetic cases, the model correctly classified 83 as non-diabetic and misclassified 17 as diabetic. Similarly, out of 54 diabetic cases, 29 were correctly identified, while 25 were incorrectly classified as non-diabetic.



```
[[83 17]
 [25 29]]
              precision    recall  f1-score   support

           0       0.77      0.83      0.80       100
           1       0.63      0.54      0.58        54

    accuracy                           0.73       154
   macro avg       0.70      0.68      0.69       154
weighted avg       0.72      0.73      0.72       154
```

The classification report offers further insights into the model's performance, presenting metrics such as precision, recall, and F1-score for each class. For non-diabetic cases, the model achieved a precision of 0.77, recall of 0.83, and an F1-score of 0.80. However, for diabetic cases, the precision was 0.63, recall was 0.54, and F1-score was 0.58. The overall accuracy of the model on the test set was 73%, indicating the proportion of correctly classified instances.



```
   Pregnancies   Glucose  BloodPressure  SkinThickness       Insulin  \
0     0.936914  1.226758  -6.953060e-01   8.087936e-16 -3.345079e-16
1     1.827813 -1.765076   2.779080e+00  -7.004289e-01 -1.254014e+00
2    -0.547919  0.010298   2.973756e-01  -2.451185e-01  5.231730e-01
3     0.936914 -0.252720  -6.953060e-01   8.087936e-16 -3.345079e-16
4    -0.547919 -1.567812   1.175571e-15   8.087936e-16 -3.345079e-16

        BMI  DiabetesPedigreeFunction       Age  Predicted_Outcome
0 -0.736094                 -0.537208  0.575118                  1
1  0.442829                 -0.564389  1.170732                  0
2  0.501048                  0.033595 -0.616111                  0
3 -0.736094                  0.785604  0.064591                  0
4  0.000000                 -1.117070 -0.956462                  0
```

The DataFrame showcasing individual predictions provides a detailed view of the model's outputs for each test instance, including their scaled feature values and the predicted outcome (0 for non-diabetic, 1 for diabetic). While the model demonstrates moderate performance with relatively high accuracy and precision for non-diabetic cases, it exhibits lower recall and F1-score for diabetic cases. This suggests potential room for improvement, perhaps through further refinement of the model or exploration of additional features to enhance its ability to correctly identify individuals at risk of diabetes.

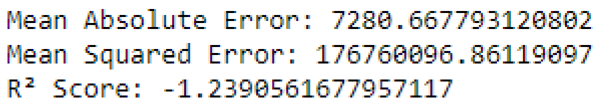
Cars Price Prediction

ALGORITHM DESCRIPTION :

For the car price prediction task, the algorithm employed is Linear Regression. Linear Regression is a simple yet effective algorithm used for modeling the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a linear equation to observed data. In this case, the model seeks to predict car prices based on features such as car length, horsepower, city and highway miles per gallon (mpg).

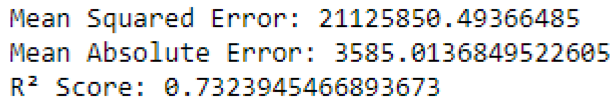
MODEL PERFORMANCE :

The model performance was initially evaluated using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 Score. These metrics provide insights into the accuracy and goodness of fit of the model.



```
Mean Absolute Error: 7280.667793120802
Mean Squared Error: 176760096.86119097
R² Score: -1.2390561677957117
```

However, the initial model exhibited poor performance, as evidenced by a negative R^2 score, indicating that the model's predictions were worse than simply using the mean of the target variable. So, I did a second analysis with only numerical values and the results were far better I still included both since I find them relevant in their own way.



```
Mean Squared Error: 21125850.49366485
Mean Absolute Error: 3585.0136849522605
R² Score: 0.7323945466893673
```

To improve the model's performance, feature selection was applied to focus on relevant features that have a significant impact on predicting car prices. Additionally, data preprocessing techniques such as one-hot encoding of categorical variables were employed to ensure compatibility with the linear regression model.

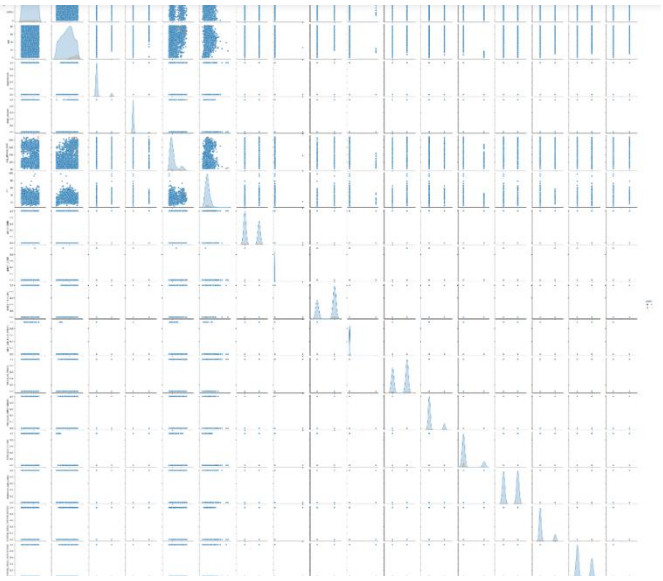
The Mean Squared Error (MSE) and Mean Absolute Error (MAE) values are substantially lower (respectively 21125850 and 3585), indicating that the model's predictions are closer to the actual values. Additionally, the R^2 Score (0.73), which measures the proportion of variance explained by the model, is substantially higher, indicating a better fit of the model to the data.

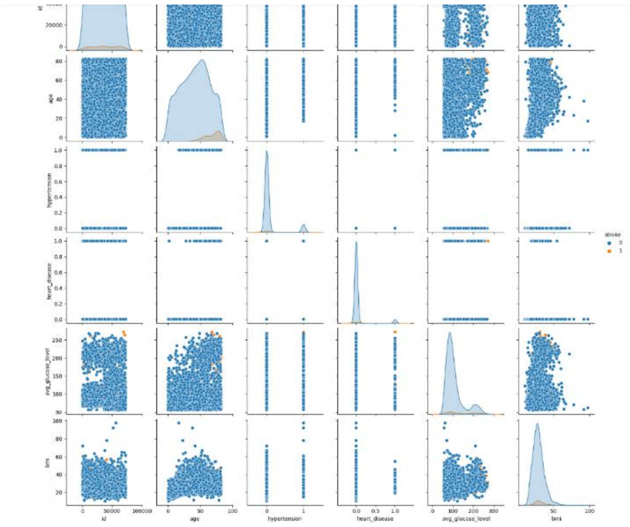
GRAPHS :

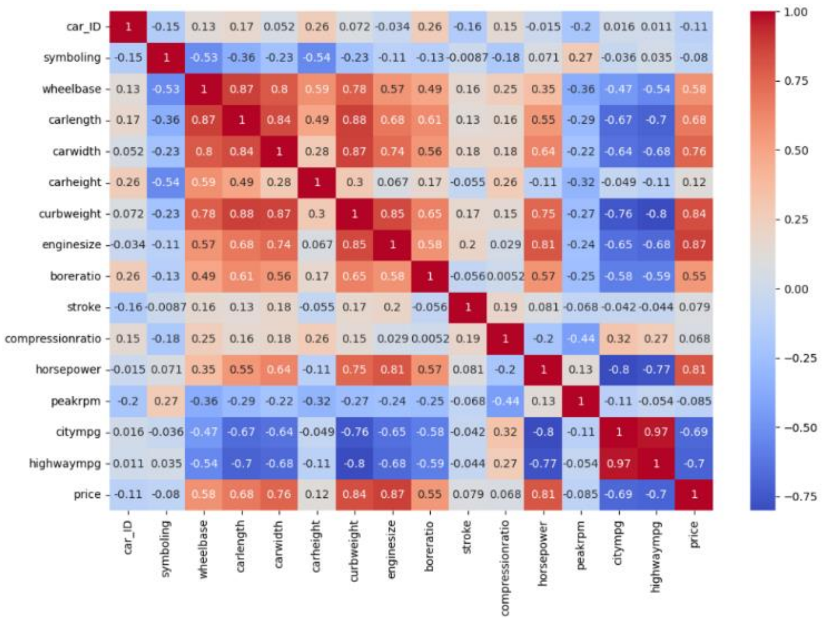
In the first version of the analysis, where only numerical features were considered, the pair plot and heatmap provided insights primarily into the relationships among these numerical features and their correlation with the target variable, car price. The pair plot allowed visualization of pairwise relationships between numerical features, aiding in identifying potential correlations and patterns.

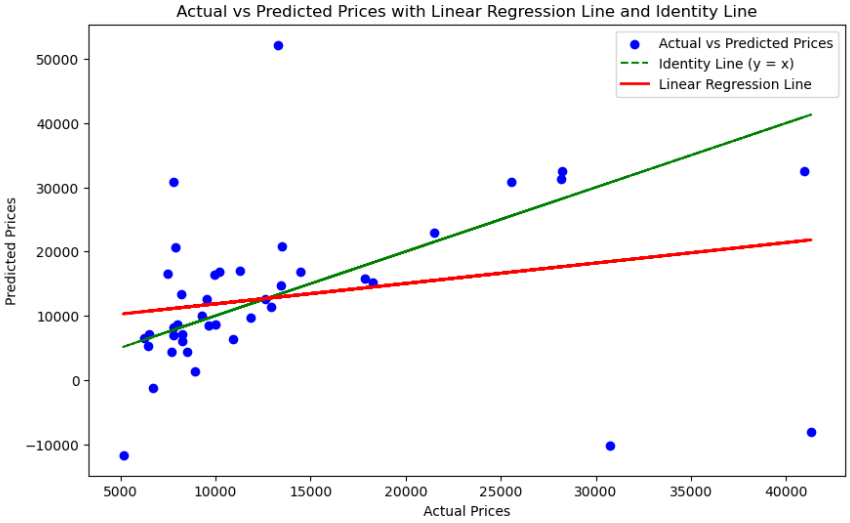
In contrast, the second version of the analysis incorporated both numerical and categorical features. The pair plot now offered a comprehensive view of relationships between all features, including both numerical and categorical ones, and their associations with the target variable. It visualized scatter plots between numerical features and the target variable, as well as box plots between categorical features and the target variable, providing insights into how different features influence car prices.

Meanwhile, the heatmap depicted correlation coefficients between numerical features, highlighting the strength and direction of their linear relationships. This analysis focused on understanding how numerical features interacted with each other and potentially influenced car prices.

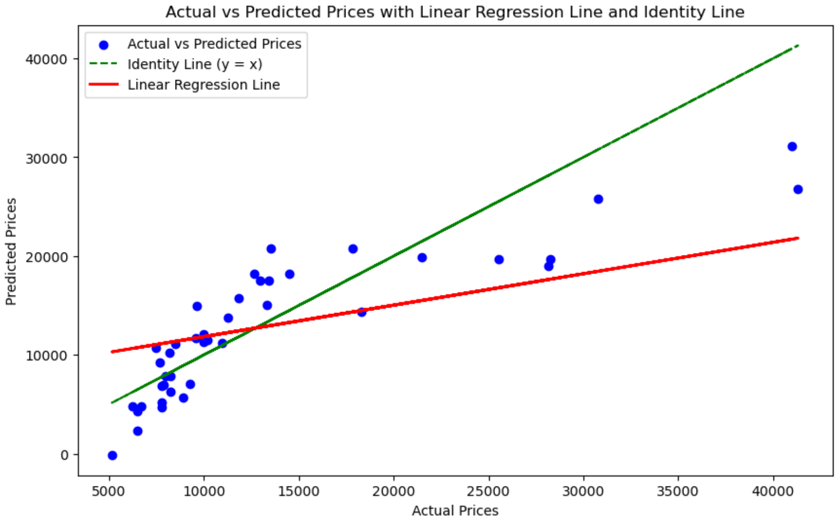


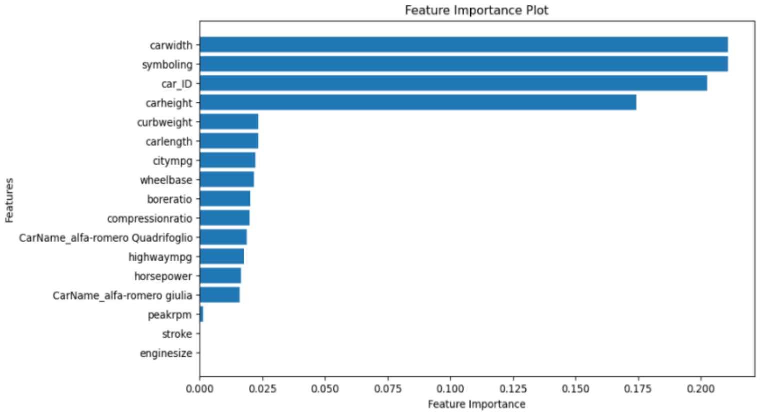






The graphical representation of actual versus predicted prices with linear regression and identity lines was intended to visually assess the model's predictive performance. However, the scatter plot suggests that the model's predictions deviate significantly from the actual prices, with a noticeable lack of alignment along the identity line even though the first analysis is closer to the y = x.

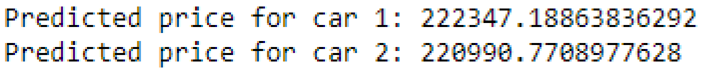




From the plot, it's evident that some features contribute more to the model's predictive performance than others. By analyzing this plot, we can identify the most influential features for the model's decision-making process.

RESULTS :

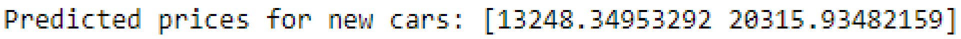
For the first with non-numerical values :



```
Predicted price for car 1: 222347.18863836292
Predicted price for car 2: 220990.7708977628
```

The results seems less accurate and too high to be precise.

With the second analysis :



```
Predicted prices for new cars: [13248.34953292 20315.93482159]
```

Those results prove how much this one even with less features is better in terms of predicting prices.

Stroke Prediction

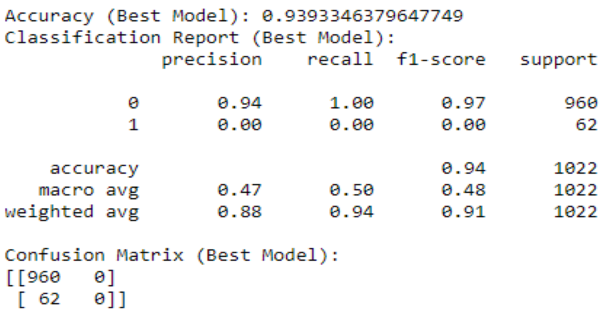
ALGORITHM DESCRIPTION :

Random Forest is an ensemble learning method used for classification and regression tasks. It operates by constructing multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of individual trees. Each tree in the forest is trained on a subset of the data, and the final prediction is based on the aggregation of predictions from all trees.

The Random Forest algorithm was employed to predict the occurrence of strokes based on various features. The algorithm works by creating a multitude of decision trees during training, each considering a random subset of features and data samples. This randomness and diversity among the trees help to mitigate overfitting and improve generalization performance.

MODEL PERFORMANCE :

The model's performance was improved through hyperparameter tuning using GridSearchCV, which exhaustively searches through a specified parameter grid to find the combination that yields the best performance (here Best Parameters: {'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}). In this case, the hyperparameters tuned included the number of estimators (trees) in the forest, maximum depth of the trees, minimum number of samples required to split an internal node, and minimum number of samples required to be a leaf node. By optimizing these hyperparameters, the model's ability to capture complex relationships within the data was enhanced, leading to improved predictive accuracy.

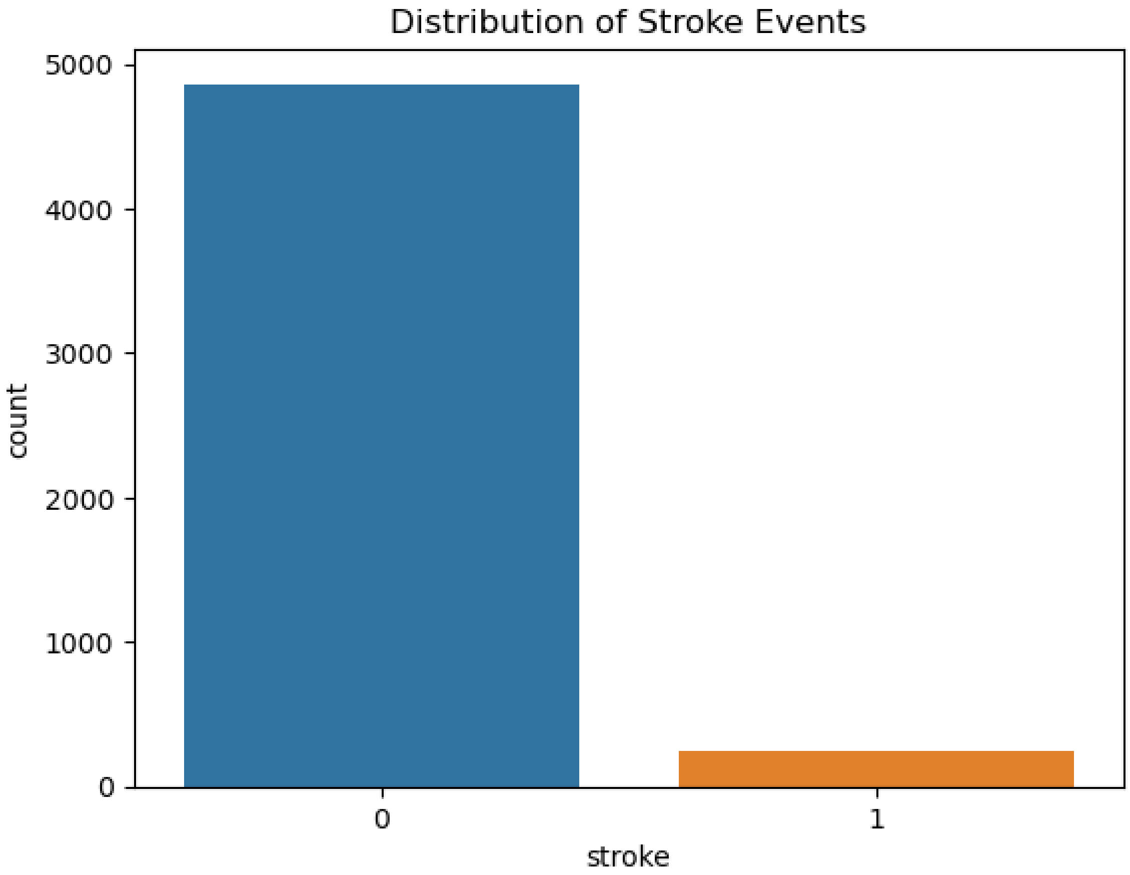


```
Accuracy (Best Model): 0.9393346379647749
Classification Report (Best Model):
              precision    recall  f1-score   support

           0       0.94      1.00      0.97       960
           1       0.00      0.00      0.00        62

    accuracy                           0.94      1022
   macro avg       0.47      0.50      0.48      1022
weighted avg       0.88      0.94      0.91      1022

Confusion Matrix (Best Model):
[[960   0]
 [ 62   0]]
```

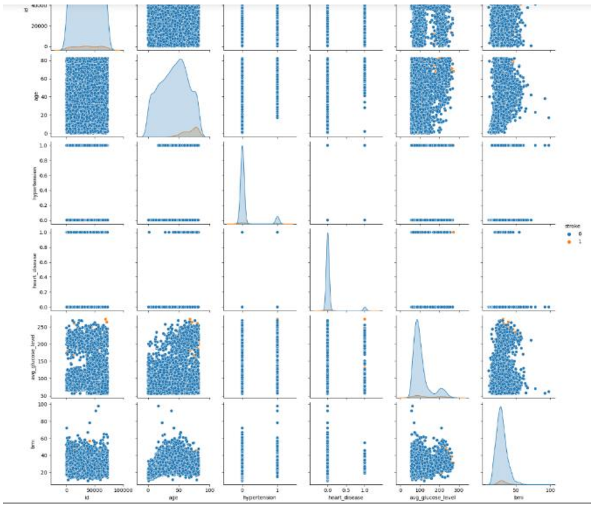
The evaluation metrics for the best model show a high overall accuracy of approximately 94%. However, a deeper analysis reveals a significant imbalance in class distribution, with class 1 having only 62 instances. As a result, the model performs poorly in terms of precision, recall, and F1-score for class 1, indicating that it struggles to correctly classify instances of that class. The confusion matrix confirms this imbalance, with all predictions falling into class 0, resulting in a lack of true positives for class 1. This highlights the importance of addressing class imbalance to improve model performance on minority classes.

GRAPHS :

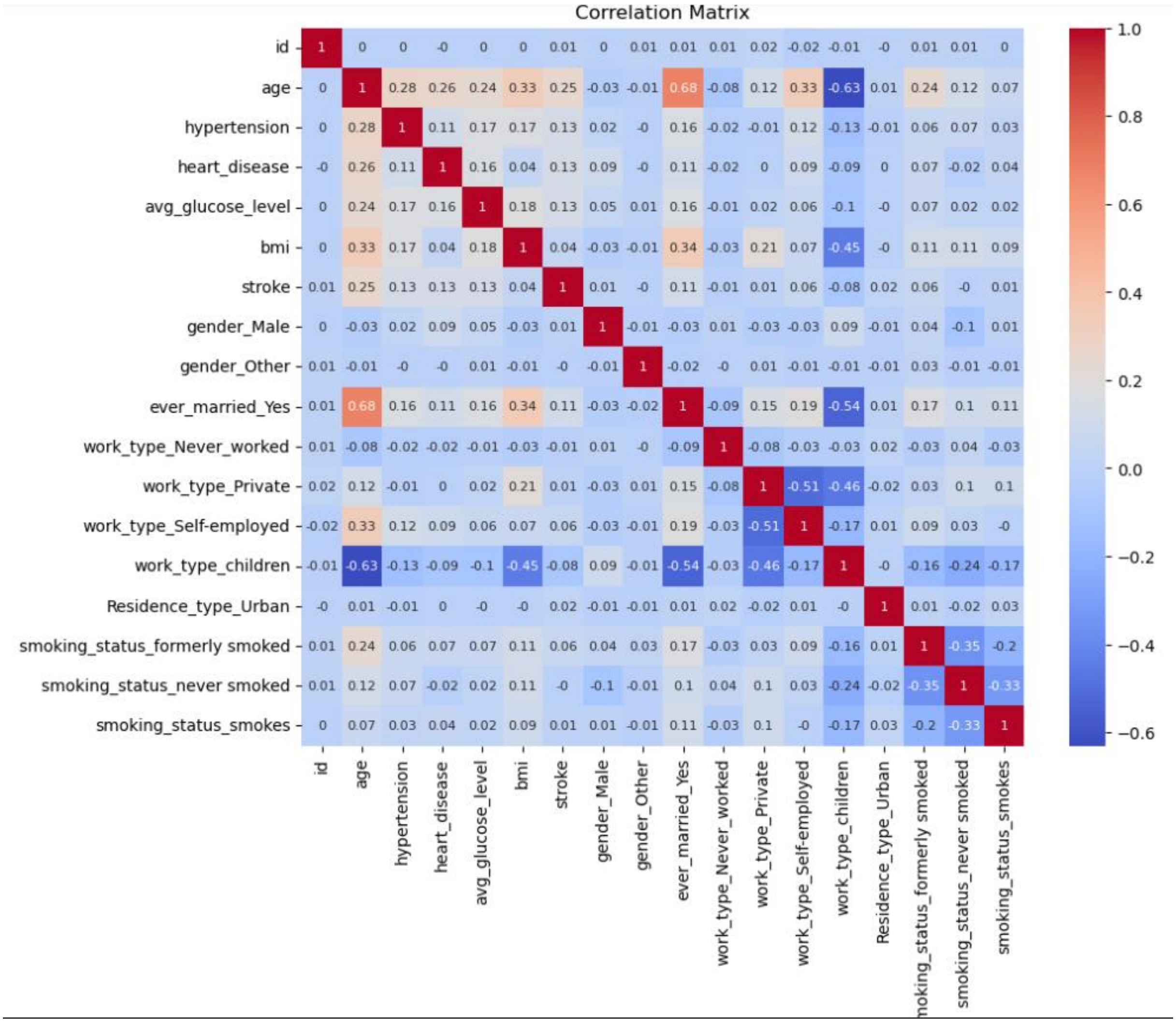
By comparing the heights of the bars, we can quickly assess the distribution of stroke events. One bar is significantly taller than the other, it indicates an imbalance in the dataset with respect to stroke events.

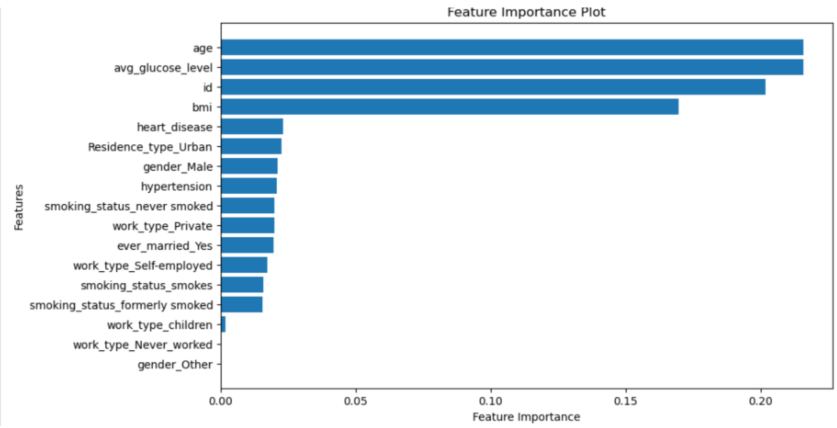


This pair plot with KDE diagonal is useful for gaining insights into potential relationships between different variables and the occurrence of stroke events in the dataset. It allows for a quick visual assessment of correlations and distributions, aiding in exploratory data analysis and feature selection for predictive modeling.



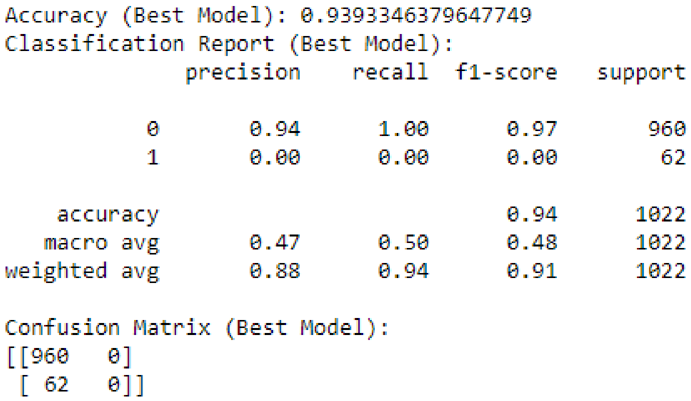
The heatmap of the correlation matrix serves as a valuable tool for identifying patterns of association between variables. It helps in understanding how variables relate to each other, which is crucial for tasks such as feature selection, multicollinearity detection, and predictive modeling.





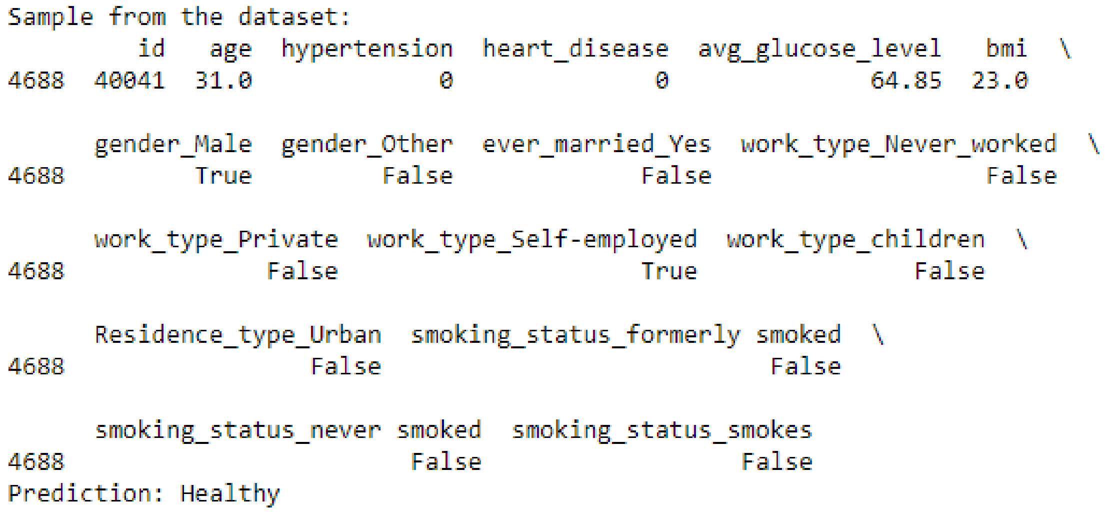
From the plot, it's evident that some features contribute more to the model's predictive performance than others. By analyzing this plot, we can identify the most influential features for the model's decision-making process.

RESULTS :



```
Accuracy (Best Model): 0.9393346379647749
Classification Report (Best Model):
              precision    recall  f1-score   support

           0       0.94      1.00      0.97       960
           1       0.00      0.00      0.00        62

    accuracy                           0.94      1022
   macro avg       0.47      0.50      0.48      1022
weighted avg       0.88      0.94      0.91      1022

Confusion Matrix (Best Model):
[[960   0]
 [ 62   0]]
```

The evaluation metrics for the best model show a high overall accuracy of approximately 94%. However, a deeper analysis reveals a significant imbalance in class distribution, with class 1 having only 62 instances. As a result, the model performs poorly in terms of precision, recall, and F1-score for class 1, indicating that it struggles to correctly classify instances of that class. The confusion matrix confirms this imbalance, with all predictions falling into class 0, resulting in a lack of true positives for class 1. This highlights some imperfection class imbalance to improve model performance on minority classes.



```
Sample from the dataset:
         id   age  hypertension  heart_disease  avg_glucose_level   bmi  \
4688  40041  31.0             0              0              64.85  23.0

      gender_Male  gender_Other  ever_married_Yes  work_type_Never_worked  \
4688         True         False             False                   False

      work_type_Private  work_type_Self-employed  work_type_children  \
4688              False                     True               False

      Residence_type_Urban  smoking_status_formerly smoked  \
4688                 False                           False

      smoking_status_never smoked  smoking_status_smokes
4688                        False                  False
Prediction: Healthy
```

The sample from the dataset represents a 31-year-old individual who appears

The provided sample from the dataset represents a 31-year-old individual with several positive health indicators. This person has no history of hypertension or heart disease, which are critical factors in assessing stroke risk. Their average glucose level is 64.85, and their BMI is 23.0, both of which fall within healthy ranges. Additionally, the individual has never smoked, is not currently married, and has been categorized as never having worked in private employment or self-employment, which may indicate a less stressful lifestyle. Living in a rural area, the person is less likely to be exposed to certain urban health risks. Based on these attributes, the model has predicted this individual to be "Healthy," indicating a low likelihood of a stroke. This prediction is consistent with their overall favorable health and lifestyle profile, supporting the model's accuracy in identifying lower stroke risks in such scenarios.