High-Dimensional Data Visualization,
Feature Extraction & Selection
Assignment 3
ID: eo9232
Name: Md Reza

Project Report

Background and Methods:

Diabetes Dataset Visualization & Feature Selection: This study proposes a machine learning model-based approach for data visualization and features selection in a higher dimensional space using PCA, *t*-SNE, UMAP, and Recursive Feature Elimination with Cross-Validation (RFECV) to identify best features to improve model performance and accuracy.

Pre-processing:

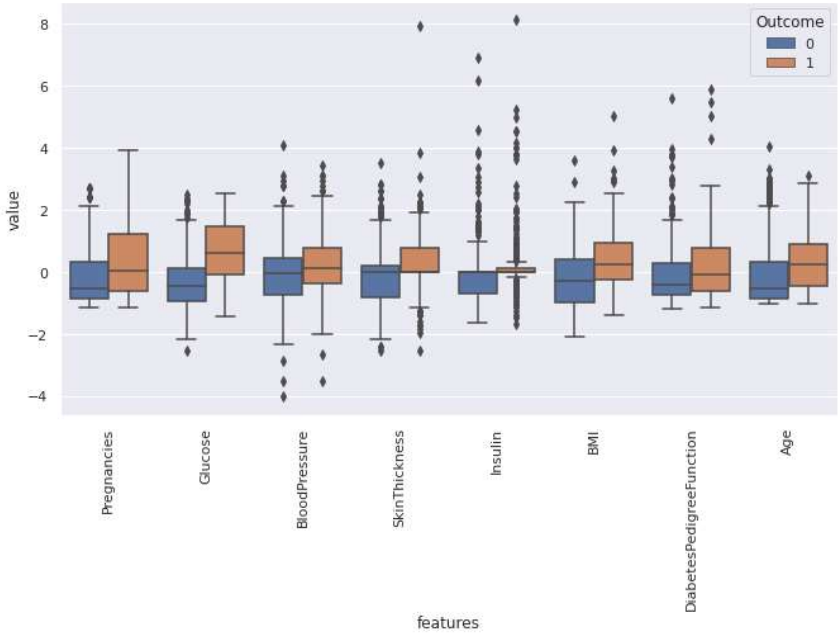
Diabetes Dataset Visualization & Feature Selection	**1.** Check for missing values **2.** Impute missing values with mean values **3.** Scaled the dataset

Hyper-Parameter Tuning:

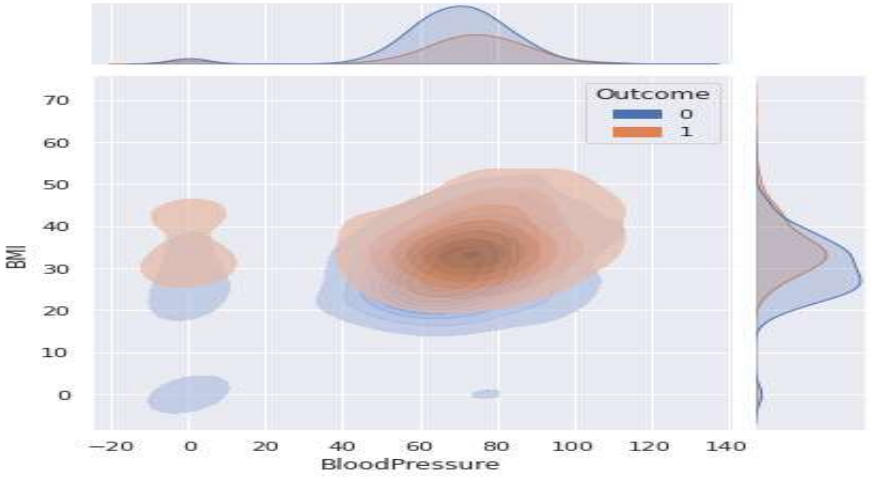
MLP Model	Grid Search with Cross-Validation	Justification
Diabetes Dataset Visualization & Feature Selection	GridSearchCV (From Scikit-learn)	To find the best parameters.

Exploratory Data Analysis & Visualization: Univariate Pair-Wise Plots & Correlation Maps are leveraged to examine the data to determine if normalization is required.

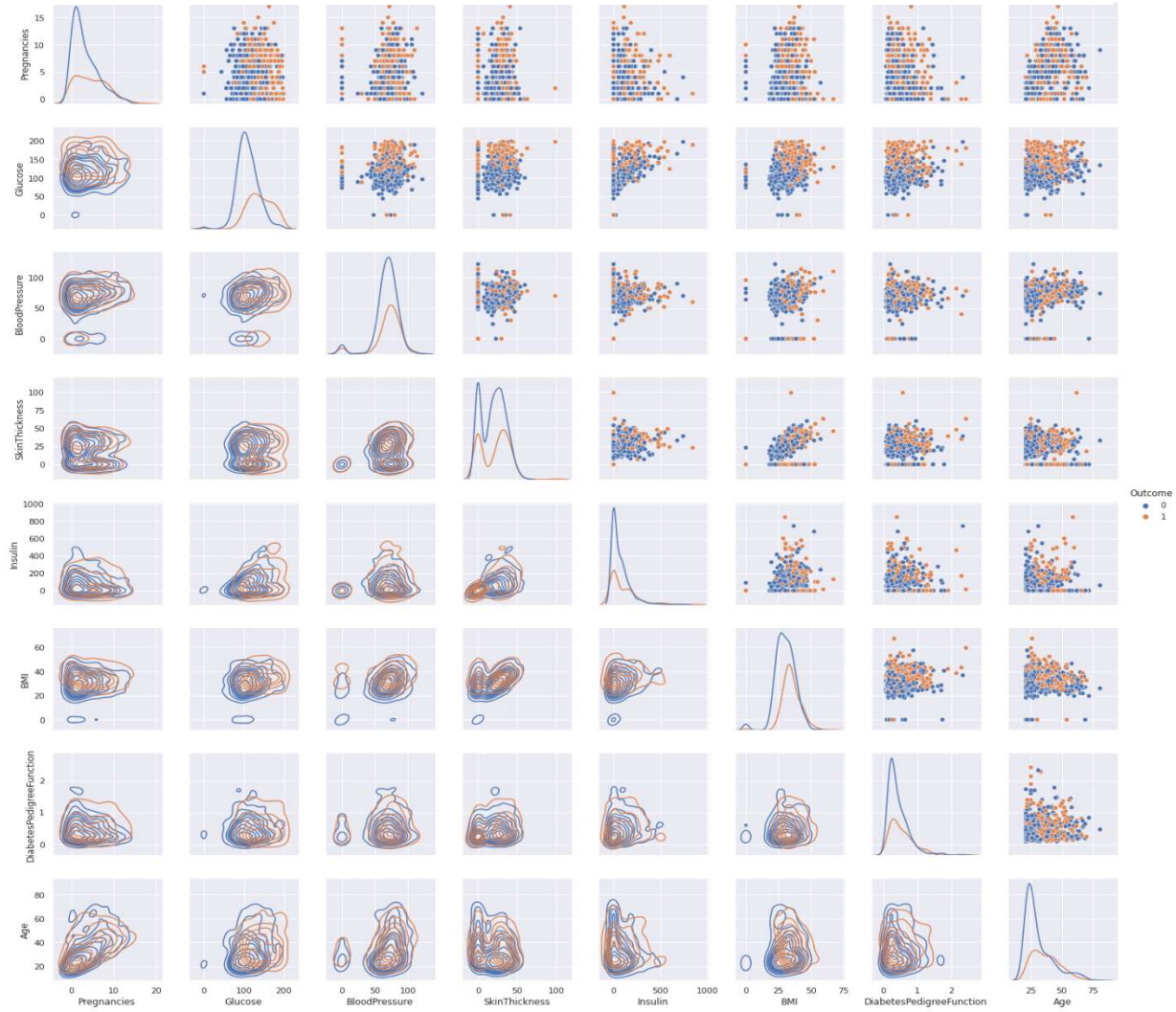
From the Boxplot on the right, it seems *Blood Pressure* and *BMI* are highly correlated. A joint plot could help us understand the correlation deeper. We aren't selecting features yet, just exploring the dataset.

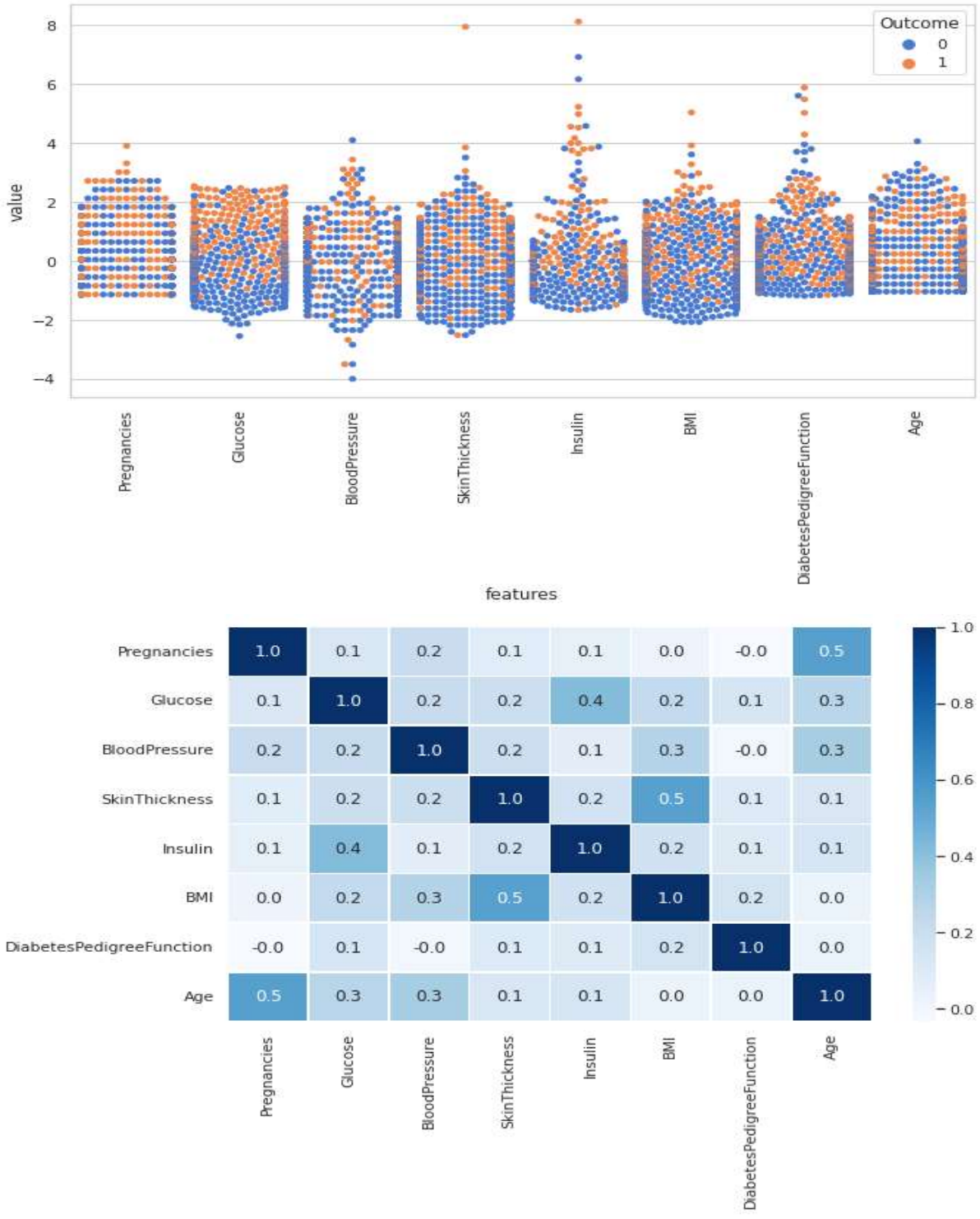


For the Joint plot on the right, it turns out that *Blood Pressure* and *BMI* are correlated. To understand the other correlations deeper, let's use a PairGrid plot. Still, we aren't selecting features yet, just exploring the dataset.



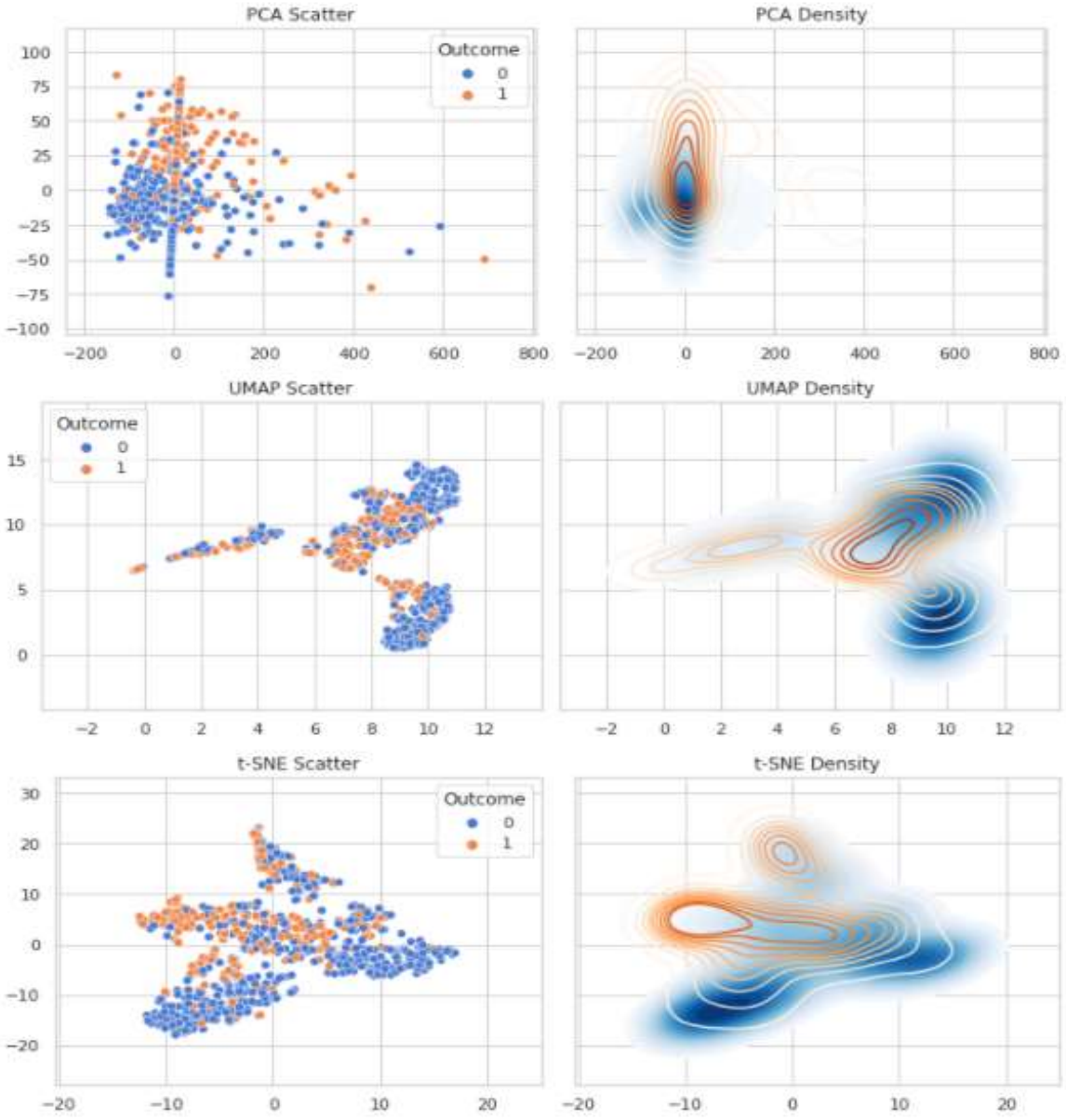
From the PairGrid plot below, it is apprehensible that all features carry some weights and possible candidates for building a better model. Probably a swarm plot could help us to learn and discover more.





From the confusion matrix above and all other correlations plots, it turned out that every feature contributes some weights towards the final model. But let us continue with the actual motivation that is feature selection & extraction, in the next preceding steps.

Visualizing High-Dimensional Data using PCA, *t*-SNE & UMAP

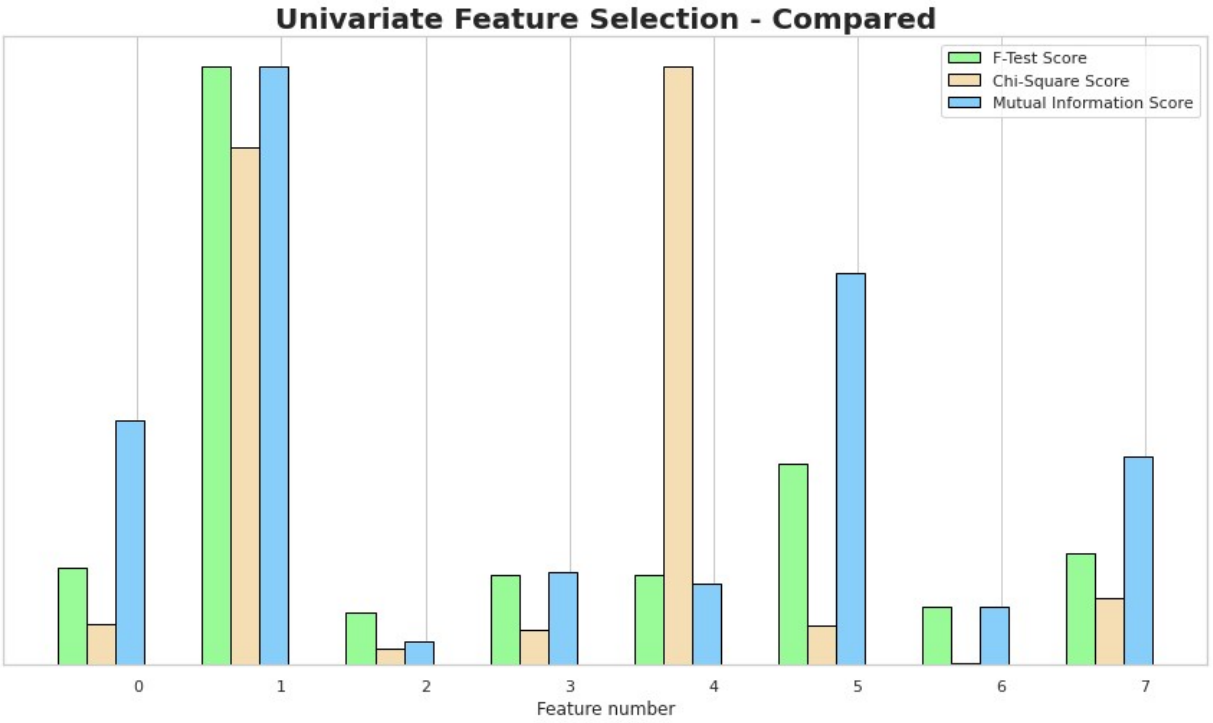


From the above plot, it turns out compared to PCA and t-SNE, UMAP performed much better to visualize the data in a higher-dimensional space.

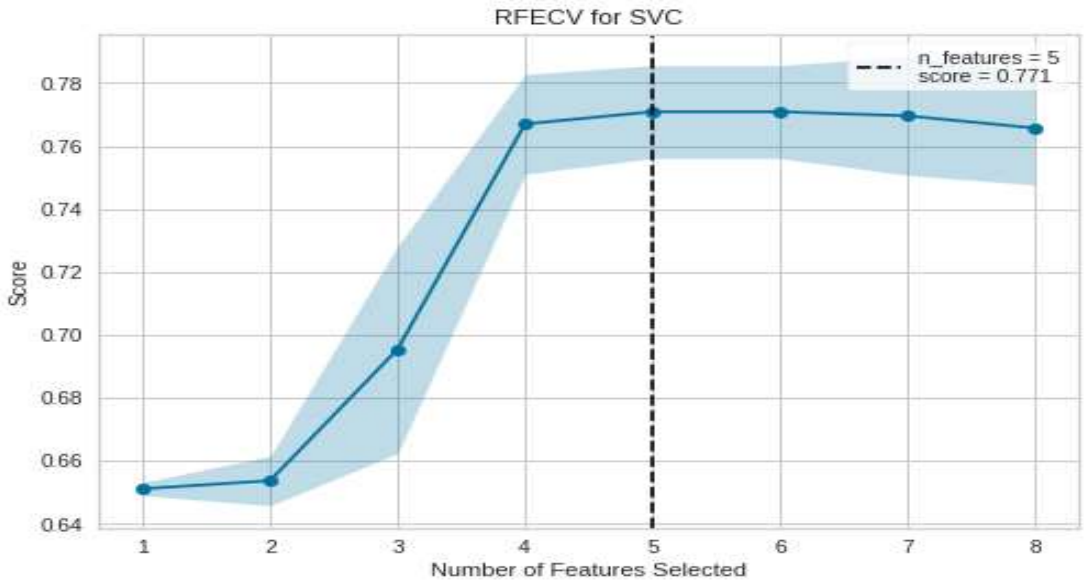
Univariate Feature Selection With:

- F-Test, Chi-Square Test, Mutual Information (Univariate Feature Selection)
- Recursive feature elimination with cross-validation (RFECV)

From the *"Univariate Feature selection – Compared"* chart below, it turns out that the Univariate Feature F-Test, Chi-Square Test, Mutual Information put some weights towards every single feature and eliminated none.



Finally, we can leverage the Recursive Feature Elimination in a Cross-Validation (RFECV) loop to find & extract the best features. With REFCV, we find the best features and discover how many features we need to build the best model. Given that MLP does not have a provision to expose "coefficient" or "feature importance" attributes, we must use an alternate classifier. I have used SVC with a linear kernel to find the best optimal parameters here. With RFECV, it turns out that the F1-Score increased and has a positive impact on the extracted optimal features. Refer to the Model Summary in the Appendix section below for details on models' performance.



APPENDIX:

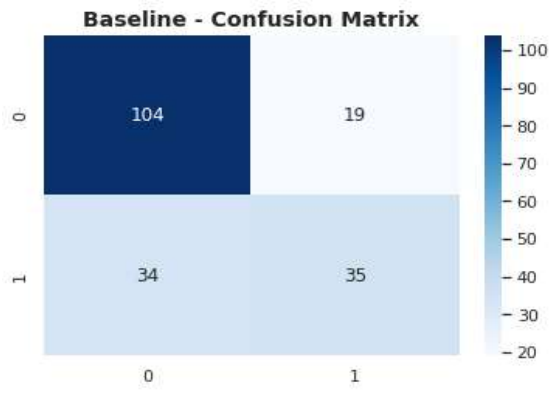
Model Summary (Univariate Feature Selection)

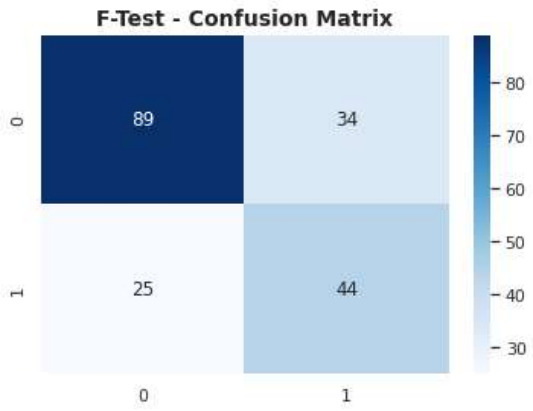
Model	Best Parameters	Precision-Recall-F1-Score-Support			
Base	`{'activation': 'tanh', 'alpha': 0.1, 'hidden_layer_sizes': (48, 24), 'learning_rate': 'constant', 'solver': 'adam'}`	0	0.75	0.85	0.80	123
		1	0.65	0.51	0.57	69
F-Test	`{'activation': 'tanh', 'alpha': 0.05, 'hidden_layer_sizes': (96, 48), 'learning_rate': 'constant', 'solver': 'adam'}`	0	0.78	0.72	0.75	123
		1	0.56	0.64	0.60	69
Chi-Square	`{'activation': 'tanh', 'alpha': 0.0001, 'hidden_layer_sizes': (96, 48), 'learning_rate': 'adaptive', 'solver': 'adam'}`	0	0.68	0.86	0.76	123
		1	0.54	0.29	0.38	69
Mutual Information	`{'activation': 'tanh', 'alpha': 0.05, 'hidden_layer_sizes': (96, 48), 'learning_rate': 'constant', 'solver': 'adam'}`	0	0.71	0.85	0.77	123
		1	0.58	0.38	0.46	69

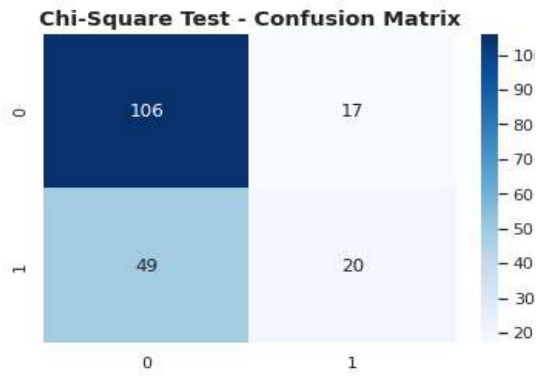
Model Summary Recursive Feature Elimination in cross-validation (RFECV)

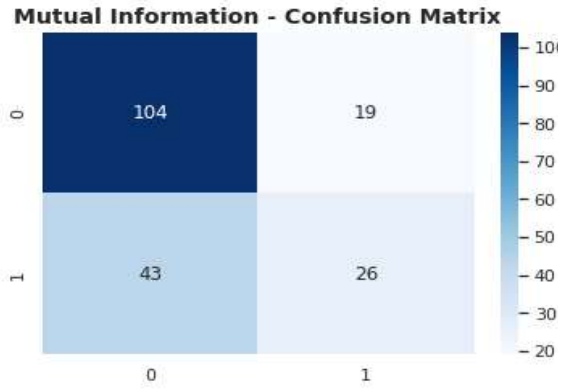
Model	Best Optimal Features	Precision-Recall-F1-Score-Support				
Recursive Feature Elimination in cross-validation (RFECV)	`['Pregnancies', 'Glucose', 'BMI','DiabetesPedigreeFunction', 'Age']`	0	0.78	0.82	0.80	123
		1	0.65	0.59	0.62	69

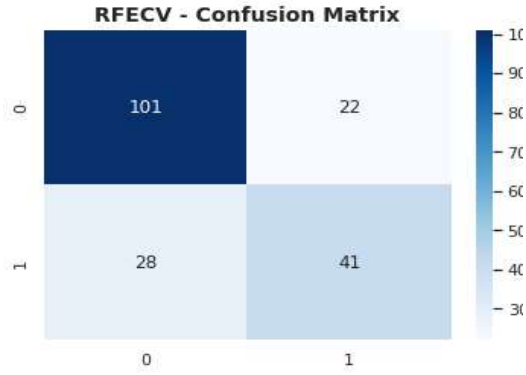
Confusion Matrix






Chi-Square Test - Confusion Matrix
106
17
49
20
0
1
0
1
90
80
70
60
50
40
30
20


Mutual Information - Confusion Matrix
104
19
43
26
0
1
0
1
90
80
70
60
50
40
30
20


RFECV - Confusion Matrix
101
22
28
41
0
1
0
1