ΕΡΓΑΣΙΑ MACHINE LEARNING

Diabetes Classification

ΟΝΟΜΑ: ΜΕΡΑΝΤΖΗΣ ΒΑΣΙΛΕΙΟΣ

Α.Μ: MTN2209

1)Εισαγωγή

Στη συγκεκριμένη εργασία χρησιμοποιήθηκε ένα σύνολο δεδομένων (dataset) διαθέσιμο στην πλατφόρμα kaggle, το οποίο περιλαμβάνει διαφόρους ιατρικούς δείκτες, οι οποίοι χρησιμοποιούνται για την ταξινόμηση (classification) δειγμάτων ως θετικά ή αρνητικά στον διαβήτη. Συνολικά το dataset περιλαμβάνει 9 χαρακτηριστικά (features). Όπως γίνεται αντιληπτό το πρόβλημα που θα αναλυθεί είναι πρόβλημα δυαδικής ταξινόμησης (binary classification), καθώς με βάση τις τιμές των εννιά διαφορετικών ιδιοτήτων ταξινομούνται δείγματα διαβήτη σε μια από τις δύο κατηγορίες (0- αρνητικό, 1- θετικό). Στον πίνακα που ακολουθεί παρουσιάζονται αυτά τα χαρακτηριστικά:

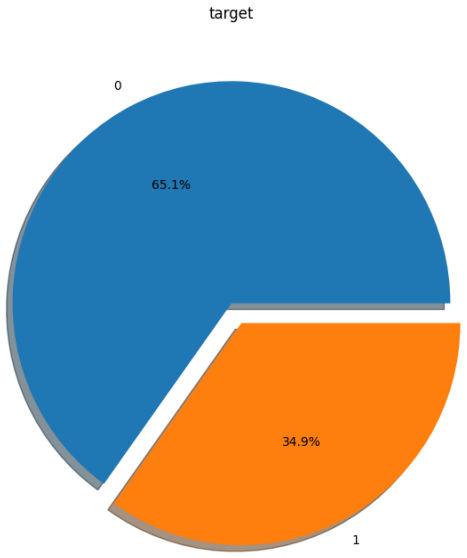
```
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
```

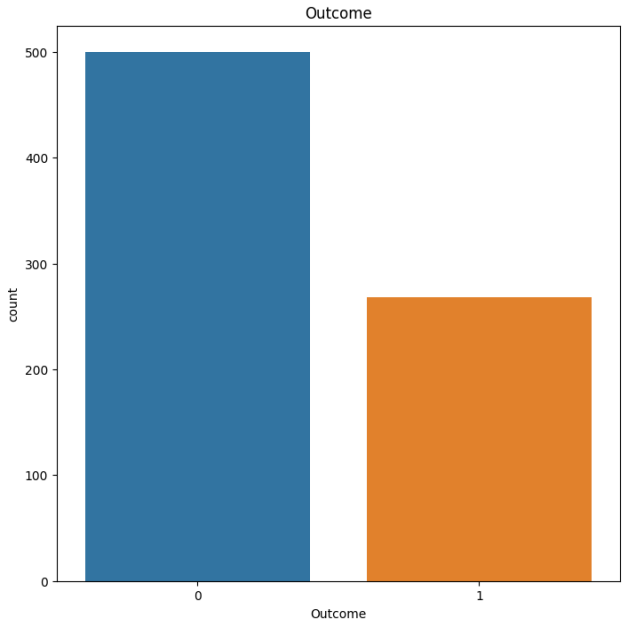
2)Basic Exploratory Data Analysis

Σε αυτό το στάδιο περιλαμβάνονται οι πίνακες και τα γραφήματα που αφορούν σε βασικά στατιστικά δεδομένα του dataset.

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

Ακολούθως φαίνεται η κατανομή των δειγμάτων με βάση το outcome ως ποσοστό και σε απόλυτους αριθμούς.

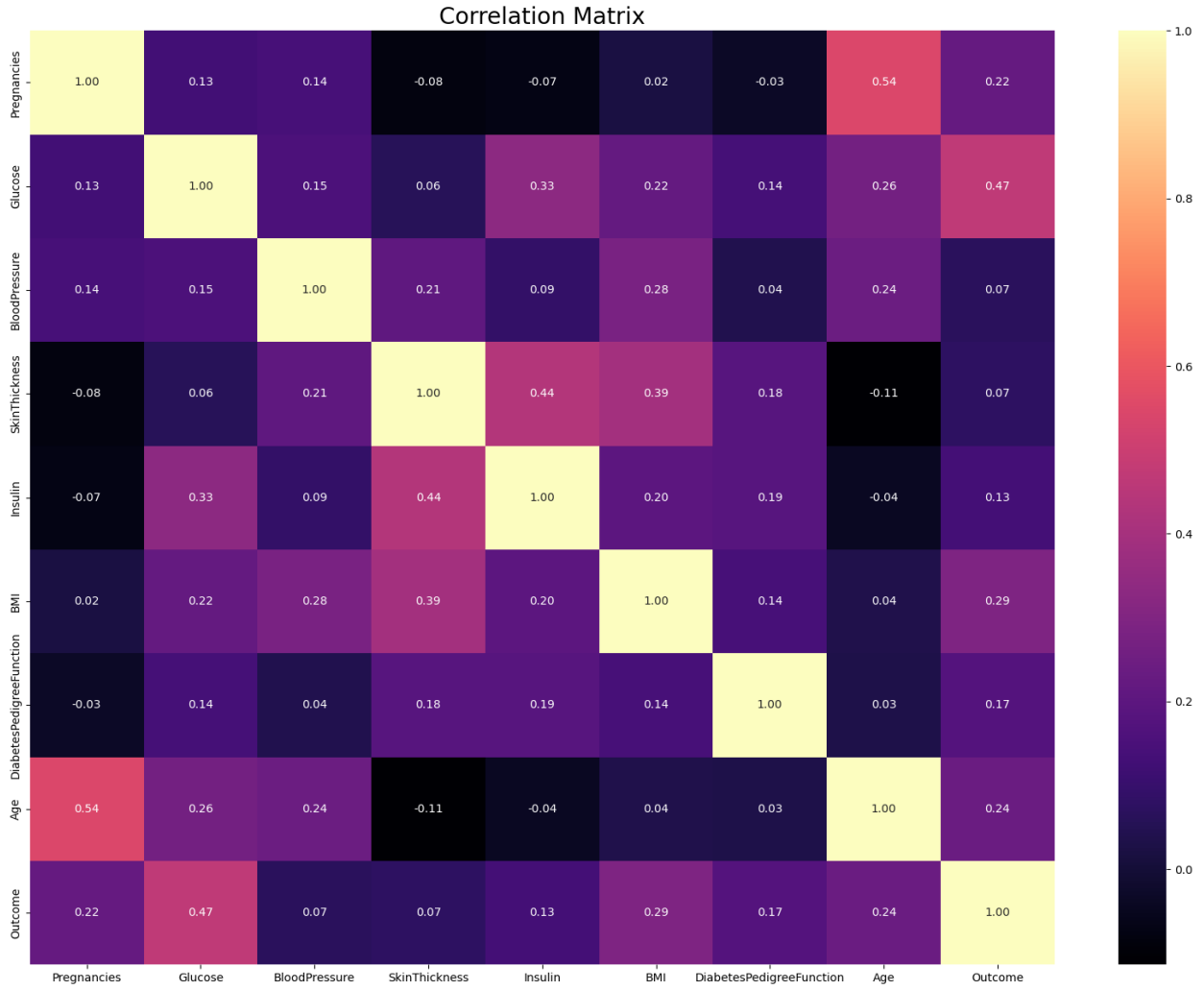




Πίνακας Συσχέτισης (Correlation matrix graph of the dataset)

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
BloodPressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
SkinThickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970	0.074752
Insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071	-0.042163	0.130548
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242	0.292695
DiabetesPedigreeFunction	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561	0.173844
Age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000	0.238356
Outcome	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356	1.000000

Heatmap


Correlation Matrix
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Outcome
1.00 0.13 0.14 -0.08 -0.07 0.02 -0.03 0.54 0.22
0.13 1.00 0.15 0.06 0.33 0.22 0.14 0.26 0.47
0.14 0.15 1.00 0.21 0.09 0.28 0.04 0.24 0.07
-0.08 0.06 0.21 1.00 0.44 0.39 0.18 -0.11 0.07
-0.07 0.33 0.09 0.44 1.00 0.20 0.19 -0.04 0.13
0.02 0.22 0.28 0.39 0.20 1.00 0.14 0.04 0.29
-0.03 0.14 0.04 0.18 0.19 0.14 1.00 0.03 0.17
0.54 0.26 0.24 -0.11 -0.04 0.04 0.03 1.00 0.24
0.22 0.47 0.07 0.07 0.13 0.29 0.17 0.24 1.00
1.0
0.8
0.6
0.4
0.2
0.0

3)Data Preprocessing

3.1) Αντιμετώπιση των Missing Values

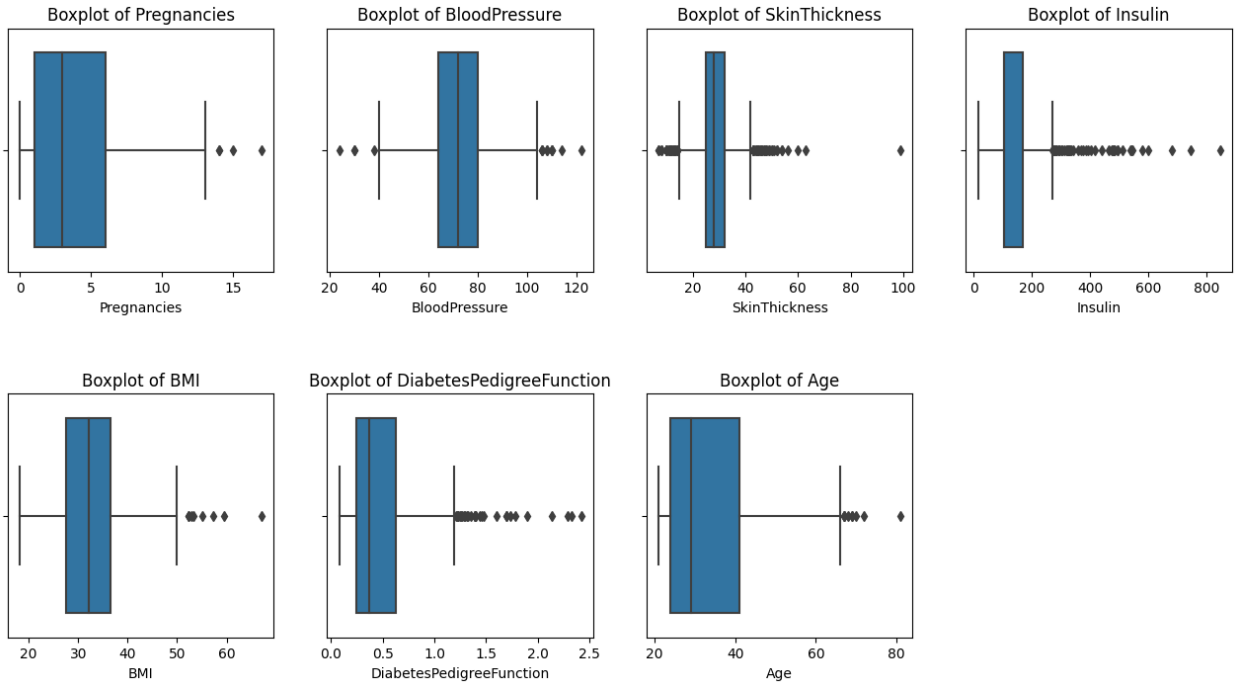
Πίνακας στοιχείων με missing values

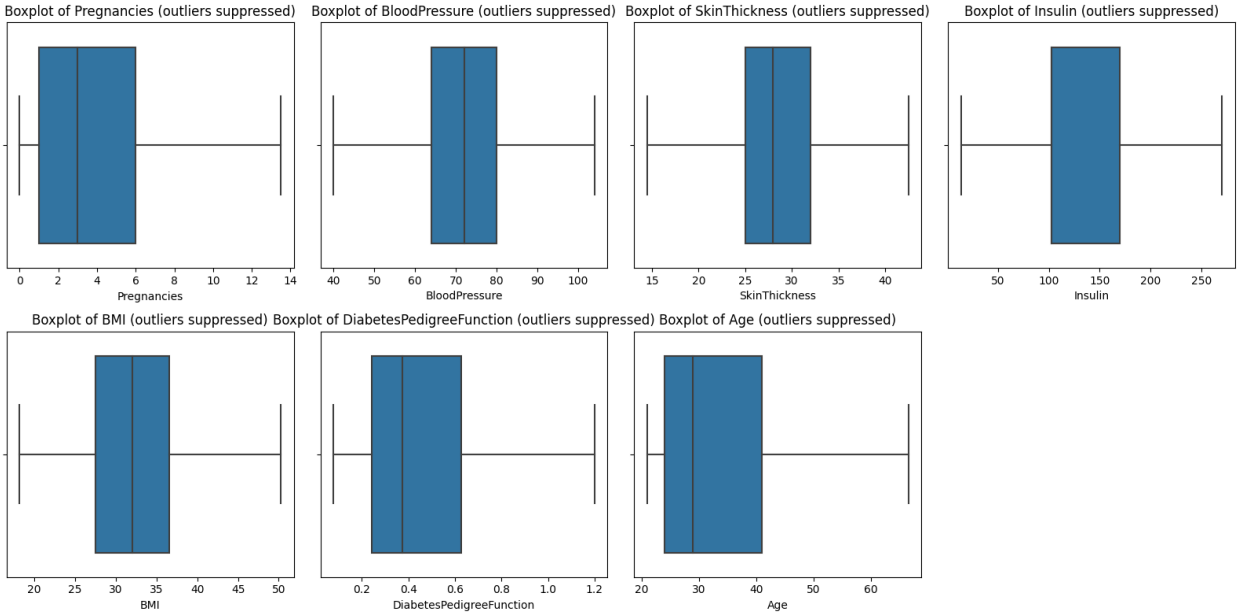
Pregnancies	0
Glucose	5
BloodPressure	35
SkinThickness	227
Insulin	374
BMI	11
DiabetesPedigreeFunction	0
Age	0
Outcome	0

Στις τιμές που λείπουν δόθηκαν η μέση τιμή των ατόμων που δεν έχουν διαβήτη (τιμή 0) και η μέση τιμή των ατόμων που έχουν διαβήτη (τιμή 1) ανάλογα με την κατηγορία στην οποία ανήκει η τιμή που λείπει και το χαρακτηριστικό στο οποίο αναφέρεται.

3.2)Αντιμετώπιση των Outliers (Outliers' Handling)

Στην συγκεκριμένη περίπτωση έγινε εντοπισμός των outliers με βάση το IQR και ύστερα έγινε suppress των τιμών των outliers στο 25% και στο 75% των τιμών των features. Παρακάτω, παρατίθενται τα boxplots πριν και αφού αντιμετωπιστούν τα outliers.





3.3) Δημιουργία Κατηγορηματικών Μεταβλητών (Categorical Variables)

Οι μεταβλητές που δημιουργήθηκαν είναι:

1) Με βάση τα επίπεδα BMI

   ["Underweight", "Normal", "Overweight", "Obesity 1", "Obesity 2", "Obesity 3"]

2) Με βάση τα επίπεδα Insulin

   ["Abnormal", "Normal"]

3) Με βάση τα επίπεδα Glucose

   ['NewGlucose_Low','NewGlucose_Normal', 'NewGlucose_Overweight']

Για να δημιουργηθούν οι κατηγορηματικές μεταβλητές χρησιμοποιήθηκε one hot encoding και συνολικά όλες οι μεταβλητές φαίνονται ακριβώς παρακάτω:

Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',

'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome', 'NewBMI_Obesity 1',

'NewBMI_Obesity 2', 'NewBMI_Obesity 3', 'NewBMI_Overweight',

'NewBMI_Underweight', 'NewInsulinScore_Normal', 'NewGlucose_Low',

'NewGlucose_Normal', 'NewGlucose_Overweight'],

dtype='object')

3.4) Standardization

Ως μέθοδος για standardization χρησιμοποιήθηκε ο RobustScaler. Ο RobustScaler είναι μια μέθοδος για scaling των μεταβλητών στη μηχανική μάθηση. Είναι παρόμοια με άλλες μεθόδους scaling, όπως ο StandardScaler ή ο MinMaxScaler, αλλά είναι πιο ανθεκτικό στα outliers. Πιο ειδικά, ο RobustScaler υπολογίζει τις παραμέτρους του scaling βασισμένος στο ενδιάμεσο εύρος τιμών (interquartile range, IQR), το οποίο είναι λιγότερο ευαίσθητο στα outliers σε σύγκριση με το μέσο και την τυπική απόκλιση. Ο RobustScaler κλιμακώνει τα χαρακτηριστικά αφαιρώντας τη μέση τιμή και στη συνέχεια διαιρώντας με το IQR. Το IQR είναι το εύρος μεταξύ του 25% (Q1) και του 75% (Q3) των δεδομένων. Αυτό εξασφαλίζει ότι το κεντρικό 50% των δεδομένων χρησιμοποιείται για την κλιμάκωση, χωρίς να λαμβάνονται υπόψη ακραίες τιμές.

4)Εκπαίδευση Μοντέλων

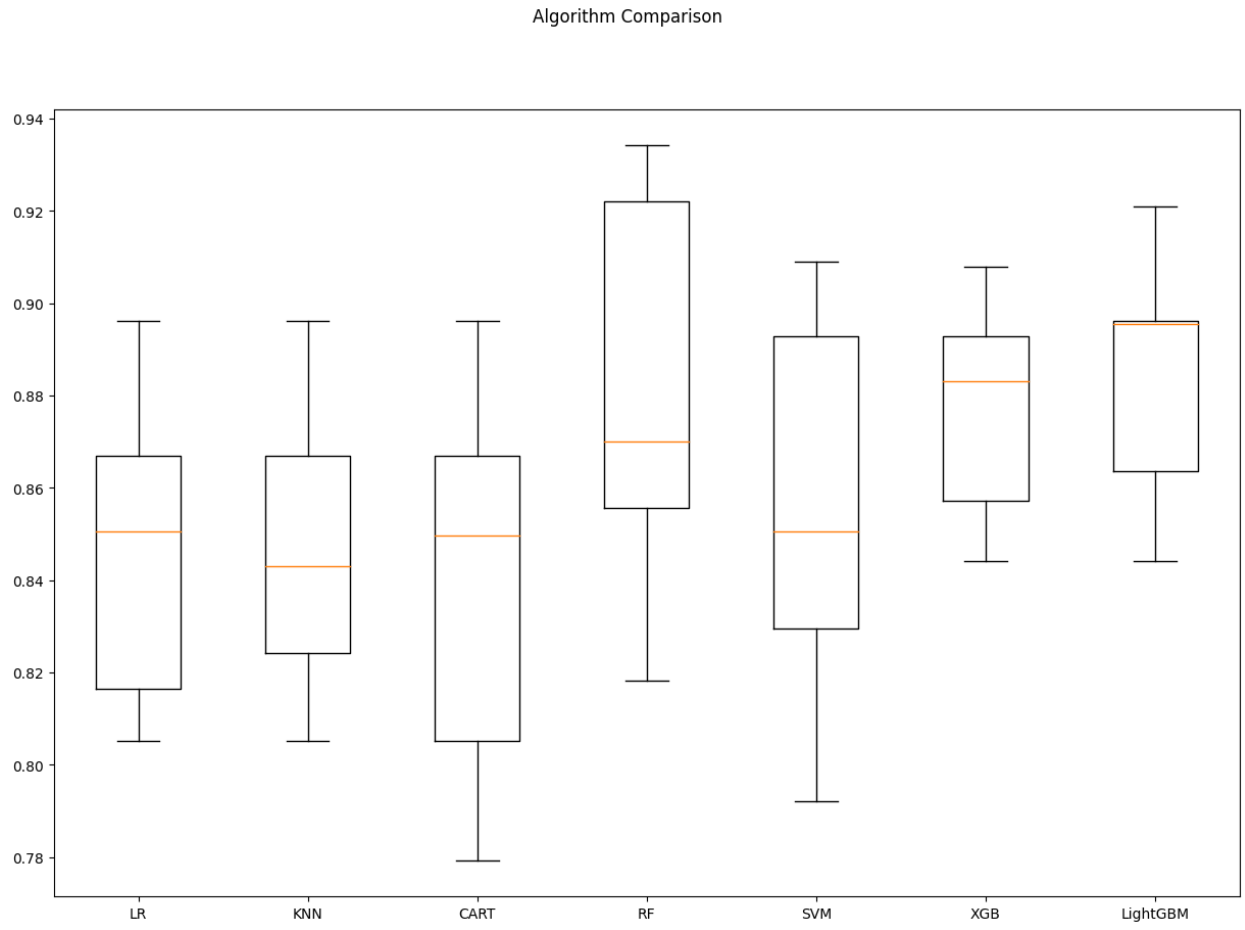
Τα μοντέλα που χρησιμοποιήθηκαν στην παρούσα εργασία είναι:

1) Logistic Regression (LR)
2) K-Nearest Neighbors (KNN)
3) Decision Tree (CART)
4) Random Forest (RF)
5) Support Vector Machine (SVM)
6) Extreme Gradient Boosting (XGBoost)
7) Light Gradient Boosting Machine (LightGBM)

Αρχικά, πραγματοποιήθηκε ένα evaluation των μοντέλων με default τιμές παραμέτρων με την τεχνική K-fold cross-validation για 10 folds. Ύστερα, πραγματοποιήθηκε fine-tuning των υπερπαραμέτρων για τα μοντέλα 4 ως 7 (RF, SVM, XGBoost, LightGBM). Τέλος πραγματοποιήθηκε k-fold cross validation για τα tuned μοντέλα.

Μέση Ακρίβεια default μοντέλων

```
LR: 0.847573 (0.031772)
KNN: 0.848941 (0.029811)
CART: 0.841217 (0.038675)
RF: 0.880246 (0.039730)
SVM: 0.855366 (0.041303)
XGB: 0.876384 (0.023083)
LightGBM: 0.884176 (0.025405)
```



Το fine-tuning πραγματοποιήθηκε με την τεχνική του GridSearch. Ακολουθούν τα set των υπερπαραμέτρων που εξετάστηκαν για κάθε μοντέλο καθώς και οι βέλτιστες παράμετροι που προέκυψαν.

1)RF (RANDOM FOREST)

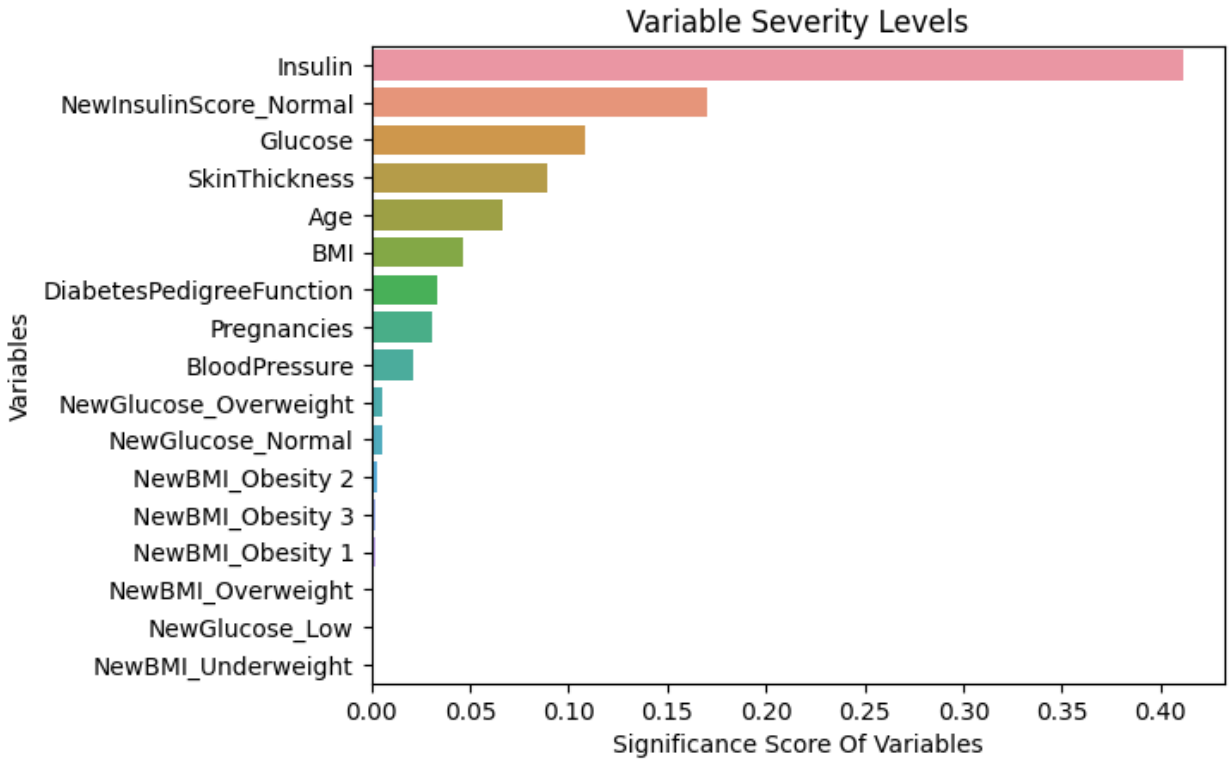
rf_params = {"n_estimators" :[100,200,500,1000],

"max_features": [3,5,7],

"min_samples_split": [2,5,10,30],

"max_depth": [3,5,8, None]}

Best RF Hyperparameters: {'max_depth': 8, 'max_features': 7, 'min_samples_split': 10, 'n_estimators': 200}



2) SVM (SUPPORT VECTOR MACHINE)

svm_params = {

"C": [0.1, 1, 10],

"kernel": ["linear", "rbf", "poly", "sigmoid"],

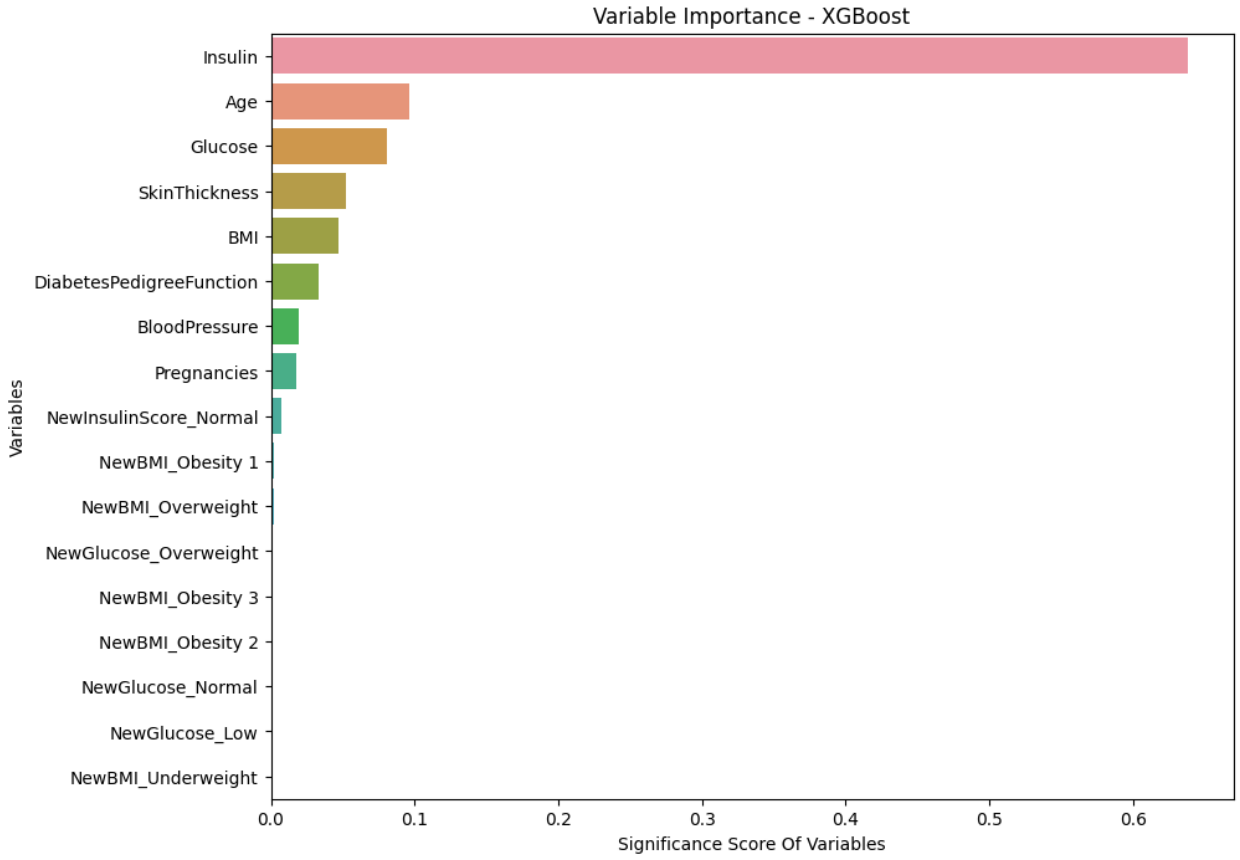
"gamma": ["scale", "auto"]}

Best SVM Hyperparameters: {'C': 1, 'gamma': 'scale', 'kernel': 'poly'}

3) XGBoost (Extreme Gradient Boosting)

```
xgb_params = {
    "n_estimators": [100, 200, 500],
    "learning_rate": [0.01, 0.1, 0.2],
    "max_depth": [3, 5, 8],
    "subsample": [0.8, 0.9, 1.0]}
```

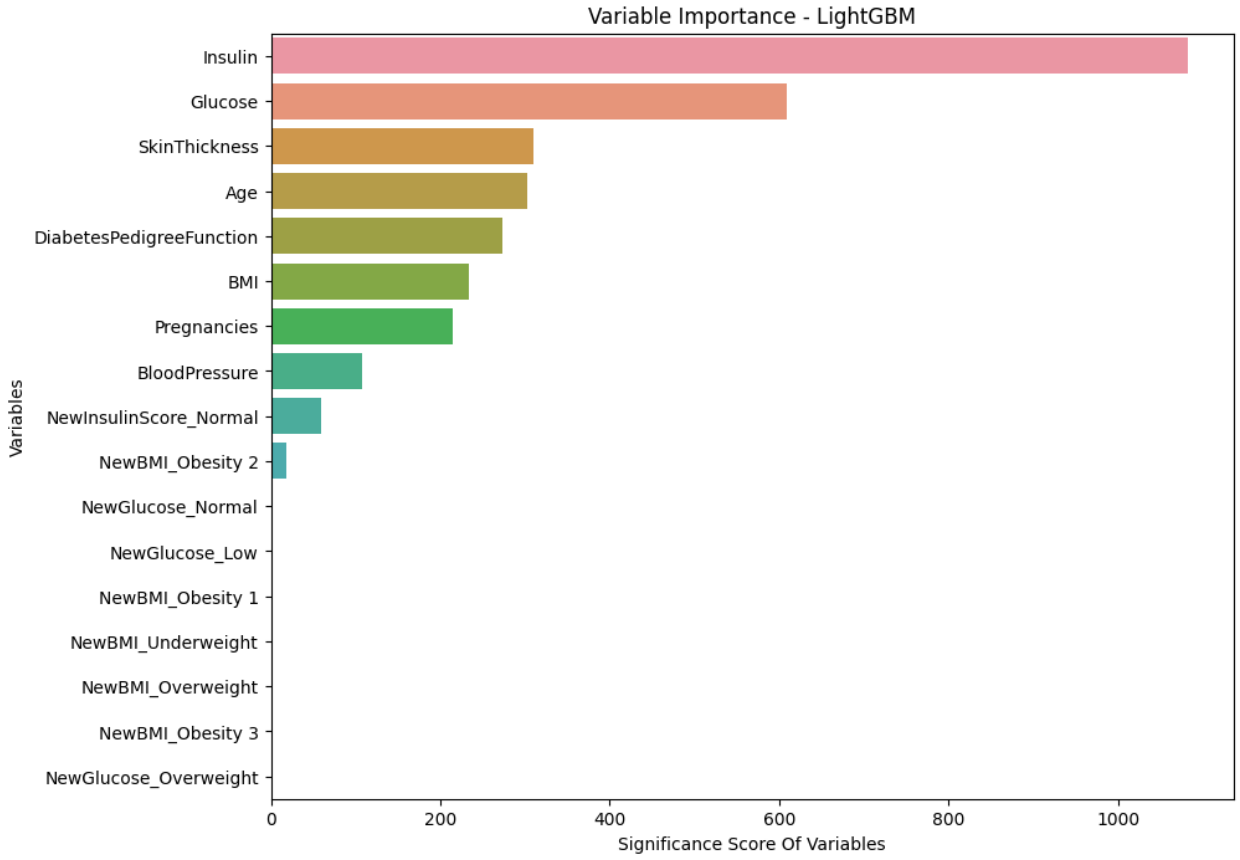
Best Gradient Boosting Hyperparameters: {'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 200, 'subsample': 0.9}



4)LightGBM (Light Gradient Boosting Machine)

```
lgbm_params = {
    "n_estimators": [100, 200, 500],
    "learning_rate": [0.01, 0.1, 0.2],
    "max_depth": [3, 5, 8],
    "subsample": [0.8, 0.9, 1.0]}
```

Best LightGBM Hyperparameters: {'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 500, 'subsample': 0.8}



Μέση Ακρίβεια tuned μοντέλων

```
Tuned RF: 0.893284 (0.029911)
Tuned SVM: 0.862013 (0.034359)
Tuned XGB: 0.897180 (0.025500)
Tuned LightGBM: 0.898462 (0.024425)
```