Εφαρμογή πρόβλεψης διαβήτη

Αθανασία-Δέσποινα Σαπουντζή και Δημήτριος Κουκουγιάννης

Εξόρυξη δεδομένων 2020-21
Εξάμηνο 8ο
Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών
Πανεπιστήμιο Θεσσαλίας, Βόλος
{asapountzi, dkoukougiannis}@e-ce.uth.gr

Περίληψη Ο διαβήτης αποτελεί σύνολο ασθενειών, το οποίο επηρεάζει τον τρόπο με τον οποίο ο ανθρώπινος οργανισμός επεξεργάζεται τη γλυκόζη του αίματος. Στην παρούσα εργασία θα προσπαθήσουμε να προβλέψουμε την ανάπτυξη διαβήτη σε ασθενείς, χρησιμοποιώντας διάφορες τεχνικές εξόρυξης δεδομένων. Το σύνολο δεδομένων που θα χρησιμοποιηθεί είναι το Pima Indians Dataset του UCI Machine learning.

Λέξεις Κλειδιά: πρόβλεψη, Python

1 Περιγραφή των δεδομένων

Το σύνολο δεδομένων προέρχεται από το National Institute of Diabetes and Digestive and Kidney Diseases. Όλοι οι ασθενείς που περιλαμβάνονται είναι γυναίκες, από 21 ετών και άνω, και ανήκουν στη φυλή Pima. Τα πεδία του dataset περιγράφονται παρακάτω:

- Αριθμός των εγκυμοσύνων μέχρι την καταχώρηση του ατόμου στο dataset
- Συγκέντρωση γλυκόζης στο πλάσμα του αίματος.
- Διαστολική πίεση του αίματος (mm Hg)
- Πάχος του δέρματος (σε mm) στους τρικέφαλους μύες
- Ινσουλίνη στον ορρό του αίματος
- Δείκτης Μάζας Σώματος
- Συνάρτηση διαβήτη γεννεαλογικού δένδρου: Η πιθανότητα εμφάνισης διαβήτη, βάσει του ιστορικού της οικογένειας
- Ηλικία (σε χρόνια)
- Αποτέλεσμα: 0 αν ο ασθενής δεν είναι διαβητικός, 1 αν είναι

1.1 Έλεγχος τιμών NULL

Αρχικά εκτελείται η παρακάτω εντολή:

```
dataset.info()
```

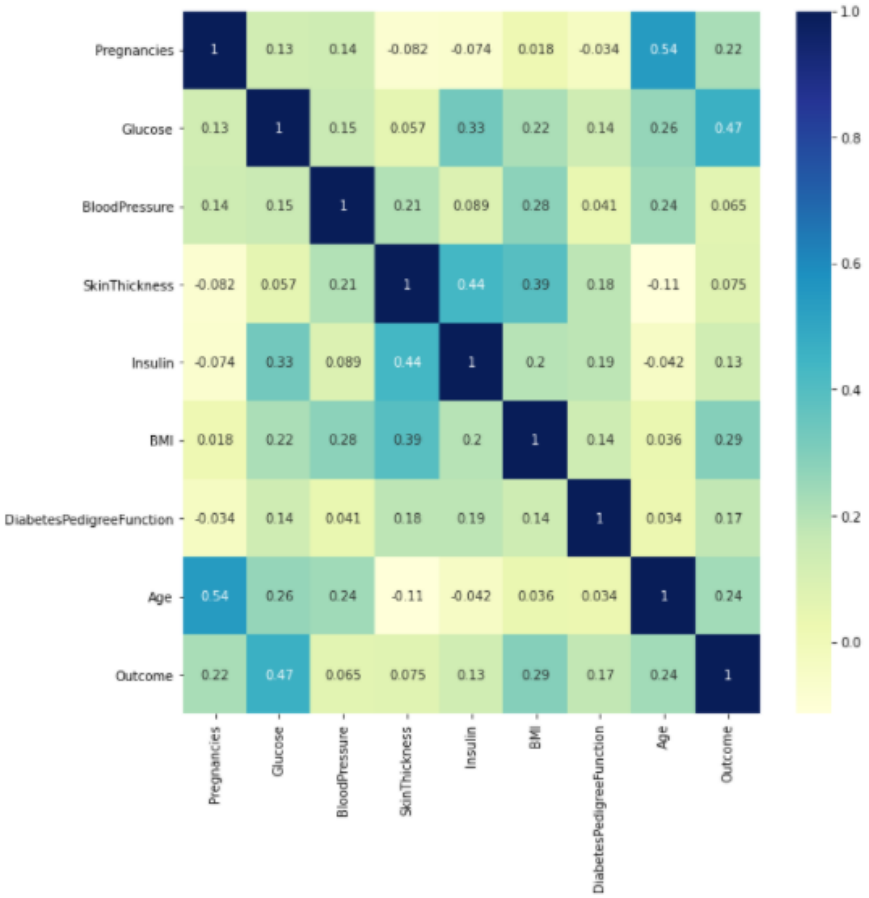
Όπως διακρίνεται στην εικόνα 1 δεν υπάρχουν τιμές NULL. Επίσης διαπιστώνουμε ότι υπάρχουν 768 καταχωρήσεις στο dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
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
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

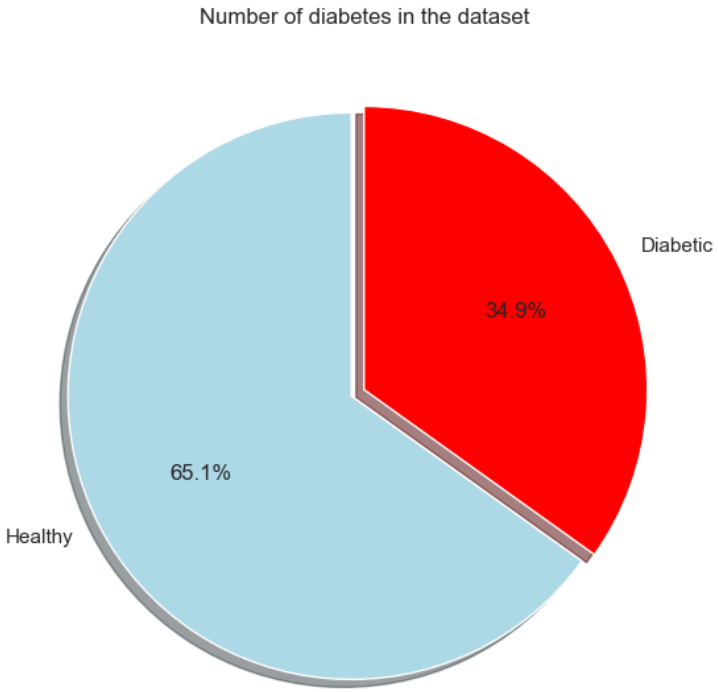
Εικ. 1. Αποτέλεσμα ελέγχου για τιμές NULL

2 Διερευνητική ανάλυση δεδομένων (EDA)

Αρχικά και αφού βρεθούν οι συσχετίσεις μεταξύ κάθε ζεύγους των γνωρισμάτων, τις απεικονίζουμε με ένα heatmap (εικόνα 2).
Επίσης, στην εικόνα 3, διακρίνεται το ποσοστό των ατόμων, οι οποίοι είναι καταχωρημένοι στο σύνολο δεδομένων. Με κόκκινο απεικονίστηκαν οι ασθενείς, οι οποίοι είχαν διαβήτη, ενώ με γαλάζιο όλοι οι υπόλοιποι.



Εικ. 2. Heatmap των συσχετίσεων των ζευγών



Εικ. 3. Ποσοστό διαβητικών και υγιών ατόμων του συνόλου δεδομένων

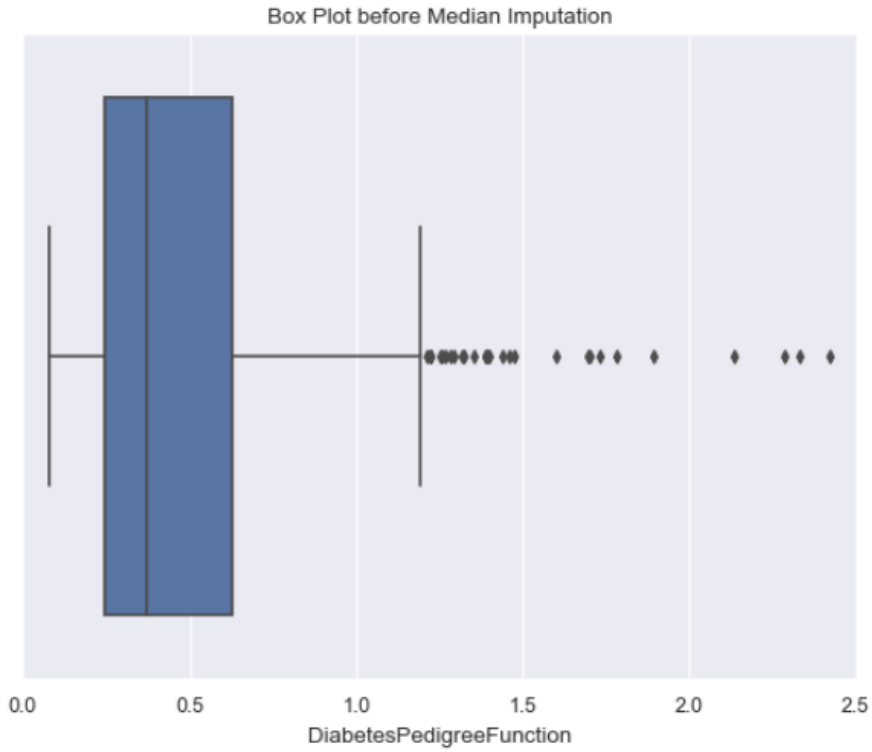
2.1 Ακραίες τιμές (outliers)

Σε αυτό το σημείο θα γίνει έλεγχος των ακραίων τιμών των γνωρισμάτων του dataset. Για την εύρεση τους χρησιμοποιείται το Interquartile range (από εδώ και πέρα IQR = $Q_3 - Q_1$). Αφού υπολογισθούν το πρώτο και τρίτο quartile (Q_1, Q_3), ο τύπος που χρησιμοποιείται είναι ο παρακάτω:

$$Lowerfence = Q_1 - 1.5(IQR)$$

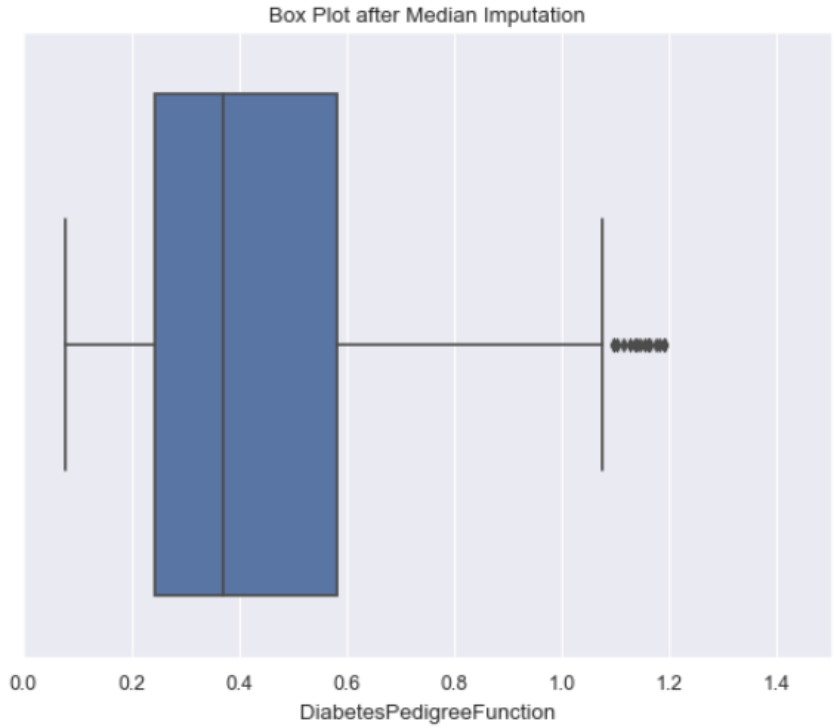
$$Upperfence = Q_3 + 1.5(IQR)$$

Στην εικόνα 4 δίνεται διακρίνεται το box plot της PedigreeFunction πριν τη διαγραφή των ακραίων τιμών, ενώ στην εικόνα 5 μετά τη διαγραφή.

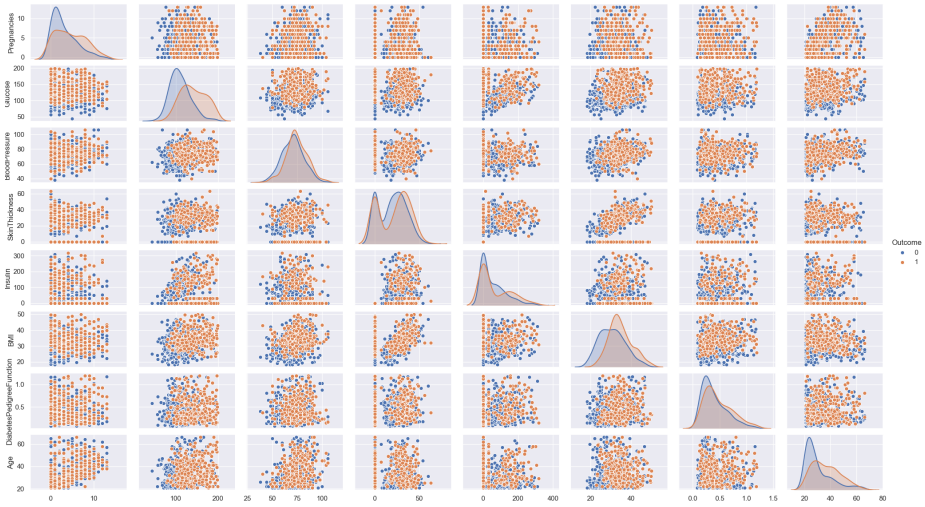


Εικ. 4. Box plot πριν τη διαγραφή των outliers

Pair plot Από το διάγραμμα συσχετίσεων της εικόνας 6 παρατηρούμε ότι δεν υπάρχουν ακραίες τιμές στο dataset.



Εικ. 5. Box plot πριν τη διαγραφή των outliers



Εικ. 6. Pair plot μεταξύ όλων των τιμών

3 Data preprocessing

Πριν ξεκινήσει η επεξεργασία των δεδομένων, επιλέγουμε ως x τις στήλες του αρχείου, εκτός από το αποτέλεσμα y (0 εάν ο ασθενής δεν είναι διαβητικός και 1 αν είναι). Στόχος μας είναι η πρόβλεψη του y, με βάση τις μεταβλητές του x.

4 Διαχωρισμός του συνόλου σε train και test set

Για να δημιουργηθεί ένα μοντέλο , το οποίο θα είναι σε θέση να προβλέψει ποιοι ασθενείς νοσούν από διαβήτη, το αρχικό σύνολο δεδομένων χωίζεται σε train set και test set, μετά την εκτέλεση τον κώδικα του αρχείου. Το train set αποτελεί το 80% του αρχικού συνόλου δεδομένων, ενώ το test set το υπόλοιπο 20%. Επίσης, στην εικόνα 7 διακρίνεται ο τρόπος με τον οποίο χωρίστηκε το αρχικό dataset, όπως περιγράφηκε.

```
Number of transactions x_train dataset:  (614, 8)
Number of  transactions y_train dataset:  (614,)
Number of transactions x_test dataset:  (154, 8)
Number of transactions y_test dataset:  (154,)
```

Εικ. 7. Ο αριθμός των transactions

5 Σύγκριση αναλύσεων

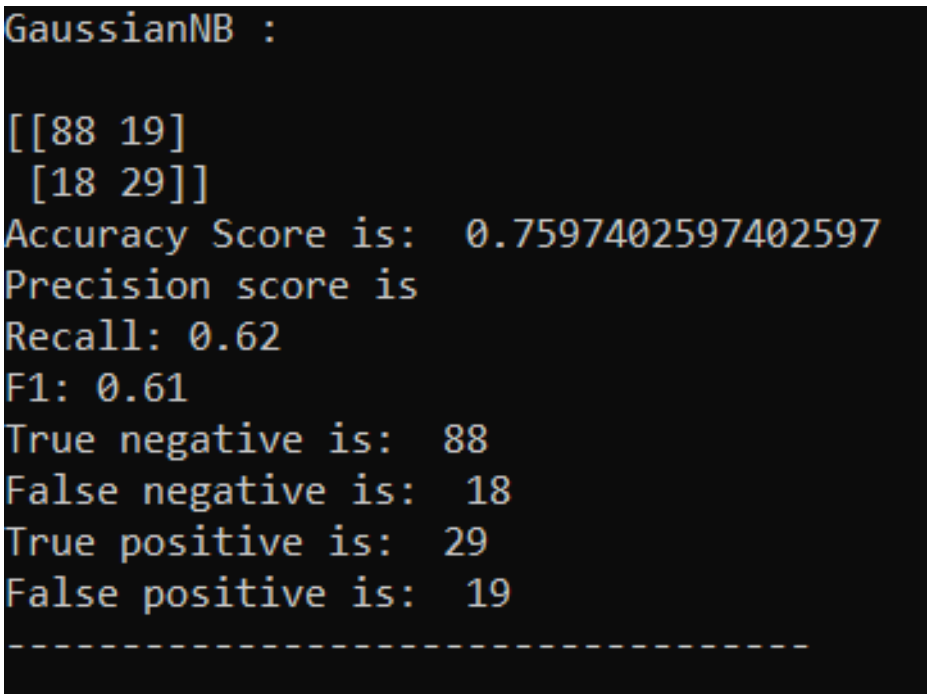
Για να επιλεχθεί η κατάλληλη ανάλυση, βάσει της οποίας θα δημιουργηθεί το μοντέλο της εφαρμογής θα συγκριθούν οι παρακάτω μέθοδοι εξόρυξης δεδομένων:

- GaussianNB
- Random Forest
- Decision tree

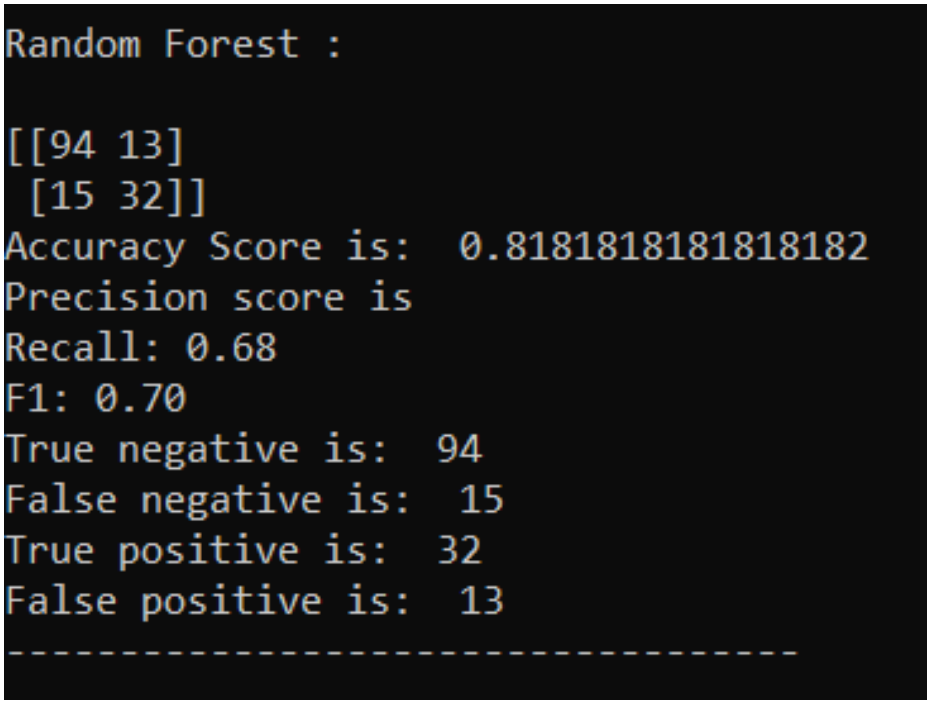
Στις εικόνες 8, 9 και 10 γίνεται απεικόνιση του confusion matrix, ο οποίος περιέχει τις τιμές true negative, false negative, true positive, false positive, και των παρακάτω μετρικές αξιολόγησεις των παραπάνω αναλύσεων:

- Accuracy
- Precision
- Recall
- F1

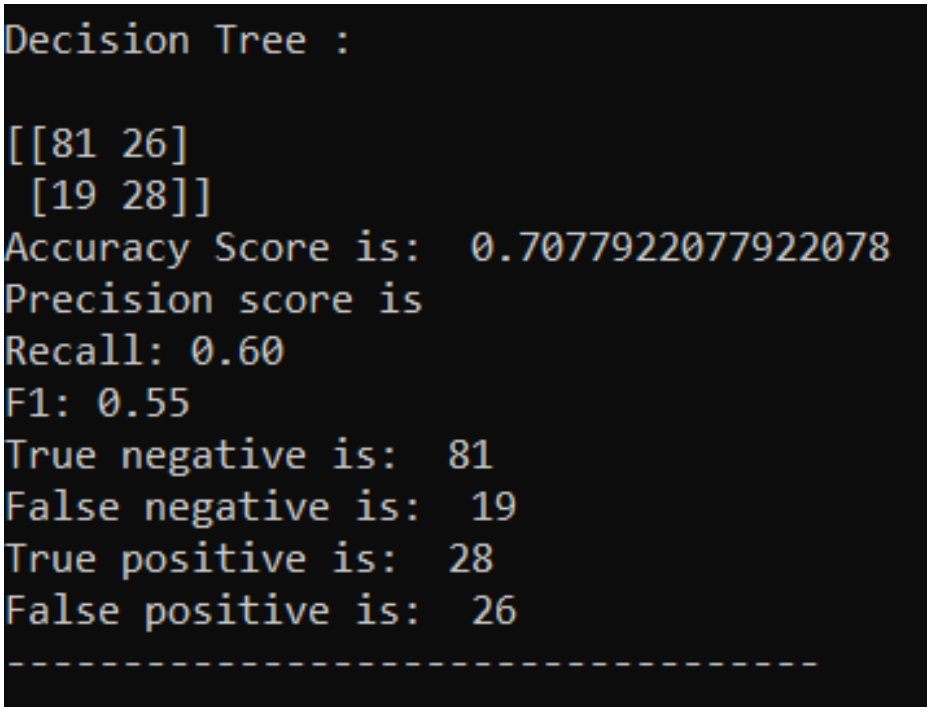
Στην εικόνα 11 διακρίνεται ο συγκριτικός πίνακας, τον οποίο λάβαμε ως αποτέλεσμα από την εκτέλεση των αναλύσεων. Διαπιστώνουμε ότι η καλύτερη μέθοδος είναι το random forest και για αυτό το λόγο θα επιλεγεί αυτή η ανάλυση για τη δημιουργία του μοντέλου της κυρίως εφαρμογής. Όπως γίνεται αντιληπτό ο Random Forest Classifier έχει τις καλύτερες μετρίκες αξιολόγησης, οπότε θα επιλεχθεί αυτή η μέθοδος εξόρυξης δεδομένων, για την ανάπτυξη της εφαρμογής.



Εικ. 8. Ανάλυση GaussianNB



Εικ. 9. Ανάλυση με Random Forest Classifier



Εικ. 10. Ανάλυση με Decision Tree Classifier

	Model	Accuracy %	Precision	Recall	F1	True negative	False negative	True positive	False positive
2	Random Forest	81.818182	0.711111	0.680851	0.695652	94	15	32	13
0	GaussianNB	75.974026	0.604167	0.617021	0.610526	88	18	29	19
1	Decision Tree	70.779221	0.518519	0.595745	0.554455	81	19	28	26

Εικ. 11. Σύγκριση των αναλύσεων

6 Δημιουργία της εφαρμογής

Όπως προαναφέρθηκε, η ανάλυση που θα χρησιμοποιηθεί από την εφαρμογή είναι το Random Forest Classifier. Αφού δημιουργηθεί το μοντέλο με βάση τη συγκεκριμένη μέθοδο, αυτό αποθηκεύεται σε ένα αρχείο. Στη συνέχεια, γίνεται έλεγχος από την εφαρμογή εάν έχουν εισαχθεί από το χρήστη όλες οι τιμές, και ανοίγεται το αρχείο και φορτώνεται το μοντέλο, έτσι ώστε να γίνει η πρόβλεψη.

7 Οδηγίες εγκατάστασης

Για την εγκατάσταση της εφαρμογής η οποία περιγράφεται στην ενότητα 7, θα πρέπει να γίνει εγκατάσταση της γλώσσας προγραμματισμού Python. Στη μπάρα αναζήτησης πατάμε Windows PowerShell, επιλέγουμε το εικόνιδιο που εμφανίστηκε και στο τερματικό του αναδυόμενου παραθύρου εισάγουμε την εντολή python. Αν η γλώσσα προγραμματισμού δεν είναι εγκατεστημένη στο μηχάνημα, τότε θα ανοίξει αυτόματα το Microsoft Store στη σελίδα της Python. Σε αυτό το σημείο θα πρέπει να πατήσουμε την επιλογή εγκατάστασης. Η Python έχει πλέον εγκατασταθεί στο σύστημα μας!

7.1 Οδηγίες εκτέλεσης

Στον φάκελο στον οποίον παραδόθηκε , βρίσκεται το αρχείο GUI.py. Για την εκτέλεση της εφαρμογής θα χρειαστεί να ανοχθεί ένα Windows PowerShell τον συγκεκριμένο κατάλογο και να εκτελεστεί η εντολή:

```
python GUI.py
```

Επειδή είναι πιθανό να χρειαστεί η εγκατάσταση επιπλέον βιβλιοθηκών, ο μεταγλωττιστής της python εμφανίζει μήνυμα λάθους. Για την εγκατάσταση της βιβλιοθήκης, η οποία χρειάζεται εγκατάσταση πρέπει να εκτελεστεί η εντολή:

```
pip install name
```

Όπου name το όνομα του πακέτου.

8 Η κυρίως εφαρμογή

Για την ανεύρεση, αν ένας ασθενής έχει διαβήτη δημιουργήθηκε μία εφαρμογή με γραφικό περιβάλλον, χρησιμοποιώντας τις βιβλιοθήκες του Tkinter, η οποία διακρίνεται στις εικόνες 12 και 13. Ο χρήστης εισάγει τα δεδομένα του ασθενή με την παρακάτω σειρά:

- Αριθμός εγκυμοσύνων (απόλυτος αριθμός)
- Γλυκόζη στο αίμα
- Διαστολική πίεση (mm Hg)
- Πάχος του δέρματος (σε mm) στους τρικέφαλους μύες
- Ινσουλίνη στο αίμα (mm U/ml)

- BMI - Δείκτης Μάζας Σώματος (απόλυτος αριθμός)
- Συνάρτηση διαβήτη γενναλογικού δένδρου (Απόλυτος αριθμός μεταξύ 0 και 1)
- Ηλικία (απόλυτος αριθμός)

Επίσης υπάρχουν δύο πλήκτρα στην κάτω μεριά της οθόνης. Με το πλήκτρο Submit, και αφού έχουν συμπληρωθεί όλες οι τιμές των πεδίων, ο χρήστης της εφαρμογής βλέπει σε ένα αναδυόμενο παράθυρο ένα από τα παρακάτω μηνύματα:

- This patient has diabetes!, εάν η πρόβλεψη είναι 1
- This patient does not have diabetes, εάν η πρόβλεψη είναι 0

Με το πλήκτρο Quit γίνεται έξοδος από το τρέχον παράθυρο και η εφαρογή σταματάει τη λειτουργία της.

Εικ. 12. Το αρχικό παράθυρο

Εικ. 13. Εμφάνιση αποτελέσματος

9 Συμπεράσματα

Στην παρούσα εργασία αναφερθήκαμε στην ανάπτυξη μίας εφαρμογής πρόβλεψης διαβήτη, για το σύνολο δεδομένων Pima Indians. Μετά τον έλεγχο διαφόρων μεθόδων διαπιστώθηκε, ότι η καλύτερη για τα συγκεκριμένα δεδομένα είναι το Random Forest Classifier, το οποίο σημείωσε ακρίβεια 81.81 %, ωστόσο και οι υπόλοιπες αναλύσεις παρουσιάζουν ικανοποιητικά αποτελέσματα.

Αναφορές

1. To dataset:
   https://www.kaggle.com/uciml/pima-indians-diabetes-database
2. EDA:
   https://www.kaggle.com/siddheshera/pima-diabetes-with-eda-12-models-beginner/notebook
3. Understanding Python pickling:
   https://www.geeksforgeeks.org/understanding-python-pickling-example/
4. Tkinter Course - Create Graphic User Interfaces in Python Tutorial:
   https://www.geeksforgeeks.org/python-gui-tkinter/
5. Python GUI - tkinter:
   https://www.geeksforgeeks.org/python-gui-tkinter/
6. Tkinter Tutorial:
   https://www.studytonight.com/tkinter
7. SKlearn:
   https://scikit-learn.org/stable/index.html
8. Εισαγωγή στην εξόρυξη δεδομένων:
   Pang-Nig Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar