ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΠΛΗΡΟΦΟΡΙΑΣ
ΤΜΗΜΑ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

Computational Methodologies & Operations Research

artificial intelligence and data analytics

Εργασία: week - (3)
Ακαδημαϊκό Έτος : 2020-2021
Διδάσκων : Καθηγητής Νικόλαος Σαμαράς

1. Titanic Data Set

Το σύνολο δεδομένων Titanic (**Titanic.xlsx**) περιλαμβάνει 2201 εγγραφές (στιγμιότυπα). Κάθε στιγμιότυπο περιλαμβάνει πληροφορίες για τους επιβάτες και το πλήρωμα του Τιτανικού όταν βυθίστηκε στις 15 Απριλίου του 1912. Κάθε στιγμιότυπο προσδιορίζεται από 4 χαρακτηριστικά:

- Το είδος της θέσης στο πλοίο (πρώτη, δεύτερη, τρίτη, πλήρωμα)
- Την ηλικία του ατόμου (ενήλικος ή παιδί)
- Το φύλο του ατόμου (άνδρας ή γυναίκα)
- Το τέταρτο χαρακτηριστικό είναι το χαρακτηριστικό εξόδου, και δηλώνει αν το άτομο που επιβίωσε διηγήθηκε την ιστορία του τραγικού συμβάντος στις αρχές.

Αναπτύξτε το προφίλ του διασωθέντος που διηγήθηκε την ιστορία χρησιμοποιώντας τεχνική κατηγοριοποίησης «δέντρο απόφασης».

Ερωτήματα.

Να γράψετε κώδικα στη γλώσσα προγραμματισμού python ο οποίος θα

(i) διαβάσει το αρχείο Titanic.xlsx.

(ii) επιλέξει ως training data set τυχαία το 80% των εγγραφών με το εξής κριτήριο: στο 80% των εγγραφών του training data set θα υπάρχει η ίδια αναλογία μεταξύ των δυο κλάσεων όπως αυτή εμφανίζεται στο σύνολο των δεδομένων. Το υπόλοιπο 20% των εγγραφών θα χρησιμοποιηθεί ως test data set.

(iii) χρησιμοποιήσει τον αλγόριθμο `DecisionTreeClassifier` για την ανάπτυξη μοντέλου πρόβλεψης όπου η δεσμευμένη μεταβλητή θα είναι η Survived και ανεξάρτητες οι υπόλοιπες μεταβλητές.

(iv) θα υπολογίζει τη μήτρα σύγχυσης (confusion matrix) για την απόδοση του προηγούμενου μοντέλου.

---

2. Cardiology Data Set

Το σύνολο δεδομένων Cardiology (**Cardiology.xlsx**) περιλαμβάνει 303 εγγραφές (στιγμιότυπα). Από αυτά τα στιγμιότυπα, 165 περιλαμβάνουν πληροφορίες για ασθενείς που δεν πάσχουν από καρδιακό νόσημα. Οι υπόλοιπες εγγραφές, 303-165=138 περιλαμβάνουν πληροφορίες για ασθενείς οι οποίοι πάσχουν από καρδιακό νόσημα. Το σύνολο δεδομένων περιλαμβάνει 13 χαρακτηριστικά εισόδου, (στήλες A ως M) και ένα χαρακτηριστικό εξόδου (στήλη N).

Στον παρακάτω πίνακα ακολουθεί περιγραφή των χαρακτηριστικών του συνόλου δεδομένων.

Attribute Name	**Values**	**Comments**
Age	Numeric	Age in years
Sex	Male, Female	Patient gender
Chest Pain Type	Angina, Abnormal Angina, No tang, Asymptomatic	NoTang=Nonanginal pain
Blood Pressure	Numeric	Resting blood pressure upon hospital admission
Cholesterol	Numeric	Serum cholesterol
Fasting Blood Sugar<120	True, False	Is fasting blood sugar less than 120?
Resting ECG	Normal, Abnormal, Hyp	Hyp=Left ventricular hypertrophy
Maximum Heart Rate	Numeric	Maximum heart rate achieved
Induced Angina?	True, False	Does the patient experience anginas a result of exercise?
Old peak	Numeric	ST depression induced by exercise relative to rest
Slope	Up, Flat, Down	Slope of the peak exercise ST segment
Numbered Colored Vessels	0, 1, 2, 3	Number of major vessels colored by fluoroscopy
Thal	Normal Fix, Rev	Normal, fixed defect,

		reversible defect
Concept Class	Healthy, Sick	Angiographic disease status

Ερωτήματα.

Να γράψετε κώδικα στη γλώσσα προγραμματισμού python ο οποίος θα

(i) διαβάσει το αρχείο Cardiology.xlsx.

(ii) επιλέξει ως training data set τυχαία το 75% των εγγραφών με το εξής κριτήριο: στο 75% των εγγραφών του training data set θα υπάρχει η ίδια αναλογία μεταξύ των δυο κλάσεων όπως αυτή εμφανίζεται στο σύνολο των δεδομένων. Το υπόλοιπο 25% των εγγραφών θα χρησιμοποιηθεί ως test data set.

(iii) χρησιμοποιήσει τον αλγόριθμο `DecisionTreeClassifier` για την ανάπτυξη μοντέλου πρόβλεψης όπου η δεσμευμένη μεταβλητή θα είναι η Class και ανεξάρτητες οι υπόλοιπες μεταβλητές.

(iv) θα υπολογίζει τη μήτρα σύγχυσης (confusion matrix) για την απόδοση του προηγούμενου μοντέλου.

HINT! Μπορείτε να χρησιμοποιήσετε τις βιβλιοθήκες pandas, numpy, sklearn, scikit-learn.

Τα ονόματα αρχείων πηγαίου κώδικα που θα παραδώσετε να είναι στη μορφή *Επίθετο_week(3-1).py* και *Επίθετο_week(3-2).py.* Επίσης, πρέπει να παραδώσετε και ένα αρχείο *Επίθετο_week(3).*docx με τις απαντήσεις στα υποερωτήματα (1-iv) και (2-iv).
