Θεσσαλονίκη, 04/02/2018

**Εργασία 1- Υπολογιστική Νοημοσύνη- Συστήματα Εμπνευσμένα από τη Βιολογία**

**Προηγμένα Συστήματα Υπολογιστών & Επικοινωνιών**

Πασιοπούλου Ιωάννα (396)

Ταμπάκη Ειρήνη- Μαρία (401)

**Α. Εισαγωγή**

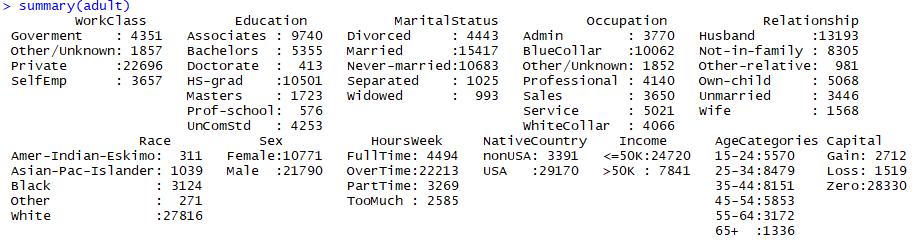
Σκοπός της εργασίας είναι να προβλέψουμε αν ένα άτομο έχει ή όχι εισόδημα μεγαλύτερο από $50.000(ΗΠΑ) ετησίως. Με βάση τα δεδομένα εκπαίδευσης για την κατασκευή ενός μοντέλου πρόβλεψης και να εφαρμόσουμε το εξαγόμενο μοντέλο στα δεδομένα δοκιμών.

**Β. Βάση Δεδομένων**

Από τις δύο επιλογές σετ δεδομένων επιλέξαμε το “Adult Dataset” από το UCI Machine Learning Repository εκτελώντας τον παρακάτω κώδικα.



Η βάση δεδομένων περιέχει δεκατέσσερις (14) ανεξάρτητες και μία εξαρτημένη μεταβλητή, που περιλαμβάνουν τόσο κατηγορικά όσο και συνεχή χαρακτηριστικά. Το σύνολο των εγγραφών είναι 32.561.



Παρακάτω ακολουθεί μια σύντομη περιγραφή των μεταβλητών της βάσης δεδομένων:

* Age: Ηλικία ατόμου. Συνεχές χαρακτηριστικό.
* WorkClass: Είδος του εργοδότη κάθε ατόμου. Κατηγορικό χαρακτηριστικό. Κατηγοριοποιείται σε: Federal-gov, Local-gov, Never-worked, Private, Self-emp-inc, Self-emp-not-inc, State-gov και Without-pay.
* Education: Επίπεδο εκπαίδευσης ατόμου. Κατηγορικό χαρακτηριστικό. Κατηγοριοποιείται σε: Preschool, 1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, Some-college, Bachelors, Masters και Doctorate.
* EducationYears: Χρόνια εκπαίδευσης κάθε ατόμου. Συνεχές χαρακτηριστικό.
* MaritalStatus: Οικογενειακή κατάσταση ατόμου. Κατηγορικό χαρακτηριστικό. Κατηγοριοποιείται σε: Divorced, Married-AF-spouse, Married-civ-spouse, Married-spouse-absent, Never-married, Separated και Widowed.
* Occupation: Τομέας απασχόλησης ατόμου. Κατηγορικό χαρακτηριστικό. Κατηγοριοποιείται σε: Adm-clerical, Armed-Forces, Craft-repair, Exec-managerial, Farming-fishing, Handlers-cleaners, Machine-op-inspct, Other-service, Priv-house-serv, Prof-specialty, Protective-serv, Sales, Tech-support, and Transport-moving.
* Relationship: Θέση ατόμου στην οικογένεια. Κατηγορικό χαρακτηριστικό. Κατηγοριοποιείται σε: Husband, Not-in-family, Other-relative, Own-child, Unmarried και Wife.
* Race: Φυλή ατόμου. Κατηγορικό χαρακτηριστικό. Κατηγοριοποιείται σε: Amer-Indian-Eskimo, Asian-Pac-Islander, Black, Other και White.
* Sex: Βιολογικό φύλο. Κατηγορικό χαρακτηριστικό. Κατηγοριοποιείται σε: Female και Male.
* CapitalGain: Κεφαλαιακά κέρδη που καταγράφονται. Συνεχές χαρακτηριστικό.
* CapitalLoss: Απώλειες κεφαλαίου που καταγράφονται. Συνεχές χαρακτηριστικό.
* HoursWeek: Ώρες εργασίας ατόμου ανά εβδομάδα. Συνεχές χαρακτηριστικό.
* NativeCountry: Χώρα προέλευσης ατόμου. Κατηγορικό χαρακτηριστικό. Κατηγοριοποιείται σε: Cambodia, Canada, China, Columbia, Cuba, Dominican-Republic, Ecuador, El-Salvador, England, France, Germany, Greece, Guatemala, Haiti, Holand-Netherlands, Honduras, Hong, Hungary, India, Iran, Ireland, Italy, Jamaica, Japan, Laos, Mexico, Nicaragua, Outlying-US(Guam-USVI-etc), Peru, Philippines, Poland, Portugal, Puerto-Rico, Scotland, South, Taiwan, Thailand, Trinadad&Tobago, United-States, Vietnam και Yugoslavia.
* Income: Το αν το άτομο έχει εισόδημα περισσότερα από $50.000(ΗΠΑ)/ έτος ή όχι. Κατηγορικό χαρακτηριστικό. Κατηγοριοποιείται σε: <= 50K και > 50K.

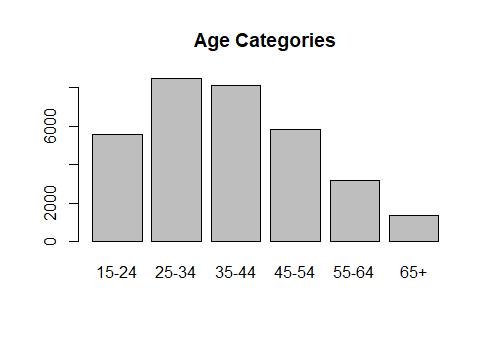
**Γ. Επεξεργασία Δεδομένων**

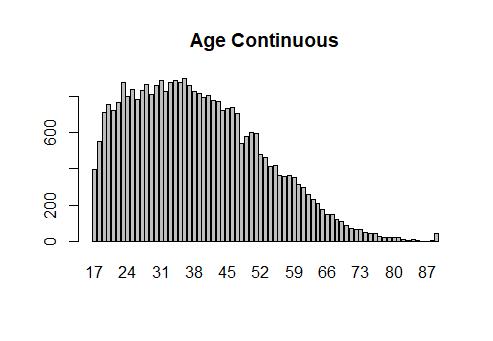
Παρατηρώντας την Βάση Δεδομένων ‘Adult’ , διαπιστώσαμε ότι είναι αναγκαία η τροποποίηση ή ακόμα και διαγραφή ορισμένων ανεξάρτητων, η χαρτογράφηση των συνεχών χαρακτηριστικών σε ποιοτικά χαρακτηριστικά και τέλος η αφαίρεση κενών πεδίων.

**Χαρτογράφηση Συνεχών Χαρακτηριστικών σε Κατηγορικά**

* **Age**

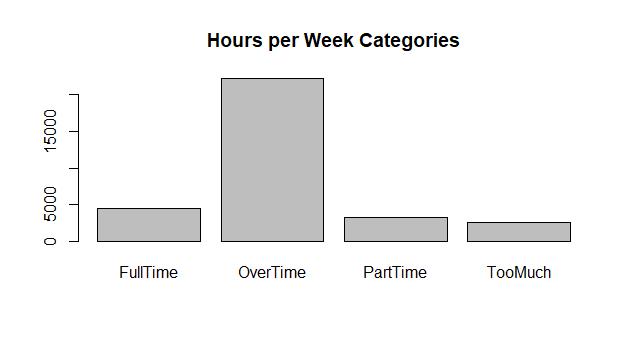
Το πεδίο Age λαμβάνει συνεχούς τύπου μεταβλητές. Διαχωρίσαμε τον πληθυσμό στις ηλικιακές ομάδες: ’15-24’, ’25-34’, ’35-44’, ’45-54’, ’55-64’ και ‘άνω των 65’.

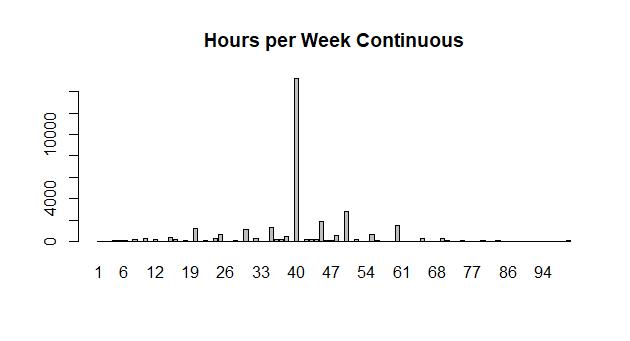




* **HoursWeek**

Ομοίως διαχωρίσαμε το πεδίο HoursWeek σε: PartTime, FullTime, OverTime και TooMuch.





* **Capital**

Τα CapitalGain και CapitalLoss αντιστοιχούν στο κέρδος ή στη ζημία του κεφαλαίου ενός ατόμου. Καταλαβαίνουμε, λοιπόν, ότι πρόκειται για συμπληρωματικά μεγέθη. Συνεπώς, για κάθε άτομο μπορούμε να έχουμε ή κέρδος ή ζημία ή τίποτα από τα δύο. Με το παραπάνω σκεπτικό, δημιουργήσαμε μία νέα μεταβλητή, την Capital, που συνοψίζει τα παραπάνω. Η μεταβλητή αυτή περιλαμβάνει διακριτές τιμές τις: Gain, Loss και Zero.

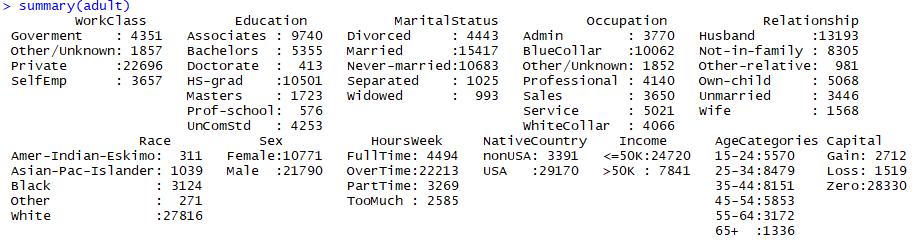
**Διαγραφή Χαρακτηριστικών**

Έπειτα, διαγράψαμε τις μεταβλητές που αντιστοιχούν στην ηλικία (συνεχείς τιμές), στα χρόνια εκπαίδευσης, επειδή στο δείγμα μας υπάρχει μεταβλητή σχετικά με το επίπεδο εκπαίδευσης και τέλος το χαρακτηριστικό Fnlwgt, καθώς προκύπτει συναρτήσει άλλων χαρακτηριστικών τους δείγματός μας. Τέλος, έχοντας φτιάξει την μεταβλητή Capital δεν μας χρειάζονται πλέον τα χαρακτηριστικά CapitalGain και CapitalLoss.

**adult$Age <- NULL   
adult$Fnlwgt <- NULL   
adult$EducationYear<-NULL  
adult$ CapitalGain <-NULL   
adult$ CapitalLoss <-NULL**

**Μείωση Κατηγοριών Ορισμένων Κατηγορικών Χαρακτηριστικών**

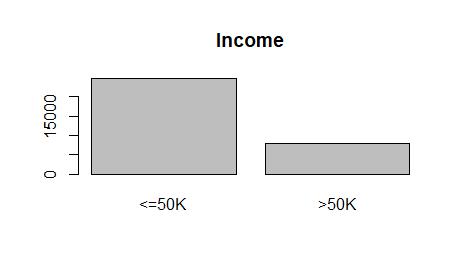
* **WorkClass**
  + Government: Federal-gov, State-gov και Local-gov.
  + SelfEmp: Self-emp-inc, Self-emp-not-inc.
  + Private: Private
  + Other/Unknown: Never-worked, Without-pay, Unknown και ‘?’
* **Education**
* UnComStd (Δεν ολοκλήρωσε τις σπουδές του): Preschool, 1st -4th, 5th -6th, 7th -8th, 9th, 10th, 11th και 12th.
* HS-grad: HS-grad
* Prof-school: Prof-school
* Associates: Assoc-acdm, Assoc-voc και Some-college
* Bachelors: Bachelors
* Masters: Masters
* Doctorate: Doctorate
* **MaritalStatus**
* Divorced: Divorced
* Married: Married-AF-spouse, Married-civ-spouse και Married-spouse-absent
* Never-married: Never-married
* Separated: Separated
* Widowed: Widowed
* **Occupation**
* Admin: Adm-clerical.
* BlueCollar: Craft-repair, Farming-fishing, Handlers-cleaners, Machine-op-inspct και Transport-moving
* WhiteCollar: Exec-managerial
* Service: Priv-house-serv, Protective-serv, Tech-support και Other-service.]
* Profecional: Prof-specialty
* Sales: Sales
* Other/Unknown: Armed-Forces και ‘?’
* **NativeCountry**
  + USA: United-States και Outlying-US(Guam-USVI-etc) [29170 εγγραφές].
  + nonUSA:Cambodia, Canada, China, Columbia, Cuba, Dominican-Republic, Ecuador, El-Salvador, England, France, Germany, Greece, Guatemala, Haiti, Holand-Netherlands, Honduras, Hong, Hungary, India, Iran, Ireland, Italy, Jamaica, Japan, Laos, Mexico, Nicaragua, Peru, Philippines, Poland, Portugal, Puerto-Rico, Scotland, South, Taiwan, Thailand, Trinadad&Tobago, Vietnam και Yugoslavia [3391]



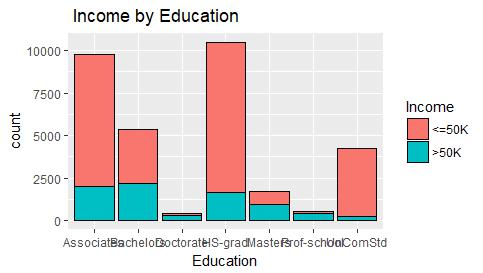
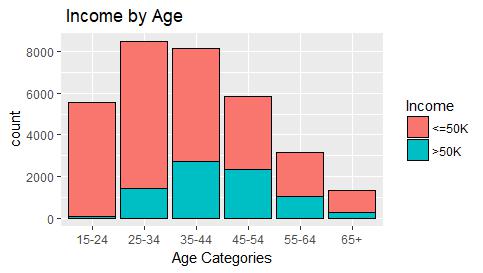
**Δ. Ανάλυση Δεδομένων**

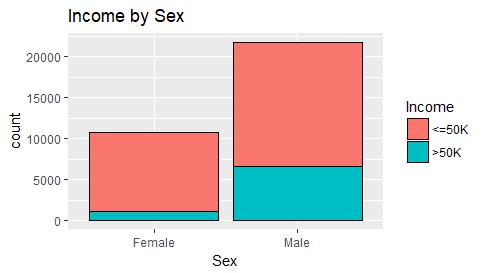
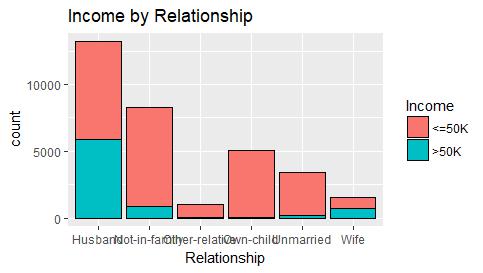
Κύριος στόχος της ανάλυσης είναι να εντοπιστούν οι παράγοντες που επηρεάζουν το ατομικό εισόδημα. Ως εκ τούτου, θα παρουσιάσουμε τη διανομή κάθε χαρακτηριστικού στο εισόδημα.

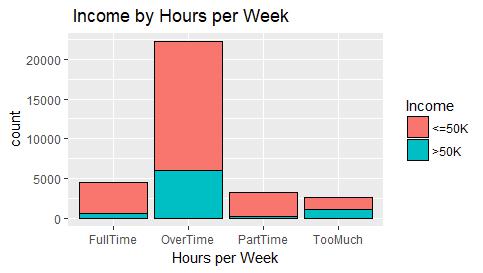
* Εισόδημα- Δύο ομάδες: εισόδημα μικρότερο ή ίσο με $50.000(ΗΠΑ) και μεγαλύτερο από $50.000(ΗΠΑ).

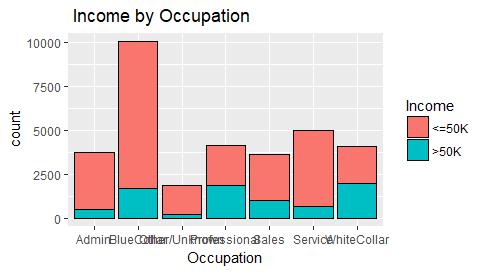


* Παρακάτω παρατίθενται τα διαγράμματα των: ηλικία, επίπεδο εκπαίδευσης, φύλο, τομέας απασχόλησης, ώρες εργασίας ατόμου ανά εβδομάδα. θέση ατόμου στην οικογένεια ως προς το εισόδημα.









**Ε. Εκπαίδευση Νευρωνικών Δικτύων**

* **Forward Propagation**

Επιλέξαμε να χρησιμοποιήσουμε το nnet που είναι ένα λογισμικό για feed-forward νευρωνικά δίκτυα, με ένα μόνο κρυφό στρώμα και για πολύ-κατηγορικά λογαριθμικά- γραμμικά μοντέλα.

Χωρίσαμε το dataframe ‘Adult’ σε training\_set και testing\_set με αναλογία 75% και 25% της βάσης δεδομένων αντίστοιχα.



Τρέξαμε το νευρωνικό μας για 300 επαναλήψεις. Στις 300 επαναλήψεις πετύχαμε βέλτιστο Accuracy. Σε περισσότερες από 300 επαναλήψεις είχαμε over fitting.



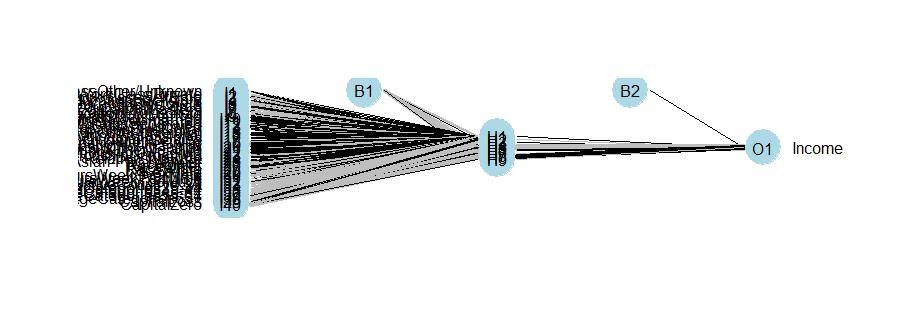
Για την εξαγωγή μετρικών από το νευρωνικό μας δίκτυο χρησιμοποιήσαμε την βιβλιοθήκη MLmetrics. Οι εντολές που χρησιμοποιήσαμε είναι οι:



Τα αποτελέσματα που λάβαμε τόσο για το training\_set, όσο και το testing\_set είναι:

|  |  |  |
| --- | --- | --- |
|  | Training Set | Testing Set |
| Accuracy | 0.8511936 | 0.8399263 |
| Precision | 0.8847496 | 0.8757977 |
| Recall | 0.9285868 | 0.9146619 |
| F-measure | 0.9061383 | 0.894808 |

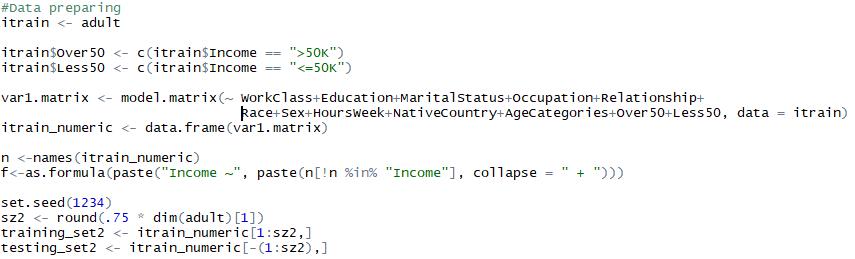
Η τοπολογία του νευρωνικού δικτύου απεικονίζεται παρακάτω:



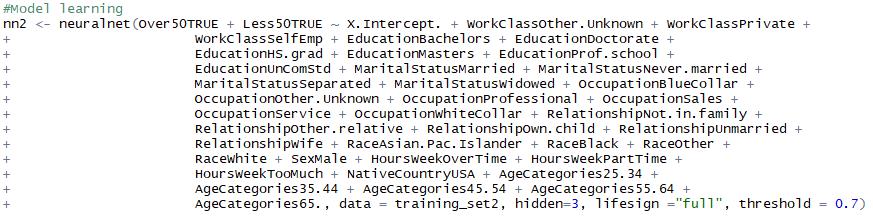
* **Back Propagation**

Επιλέξαμε να χρησιμοποιήσουμε το neuralnet που είναι ένα λογισμικό για back propagation.

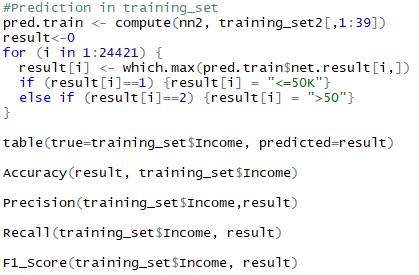
Χωρίσαμε το dataframe ‘Adult’ σε training\_set και testing\_set με αναλογία 75% και 25% της βάσης δεδομένων αντίστοιχα.

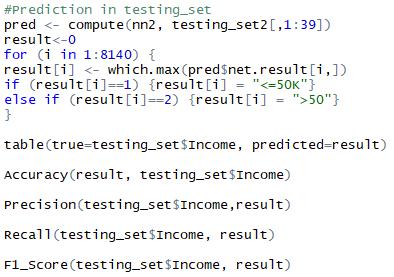


Τρέξαμε το νευρωνικό μας για threshold 0.16, αλλά δεν κατάφερε να φτάσει κάτω από 0,7.



Για την εξαγωγή μετρικών από το νευρωνικό μας δίκτυο χρησιμοποιήσαμε την βιβλιοθήκη MLmetrics. Οι εντολές που χρησιμοποιήσαμε είναι οι:





Τα αποτελέσματα που λάβαμε τόσο για το training\_set, όσο και το testing\_set είναι:

|  |  |  |
| --- | --- | --- |
|  | Training Set | Testing Set |
| Accuracy | 0.7052536751 | 0.6966830467 |
| Precision | 0.8677448609 | 0.8684532925 |
| Recall | 0.9275635502 | 0.9218140442 |
| F-measure | 0.8966576426 | 0.8943384324 |

Η τοπολογία του νευρωνικού δικτύου απεικονίζεται παρακάτω:

