# Decision Trees

PhDc Chasapi Maria Konstantina

PhDc Chasapi Lamprini

# Δημιουργία Δέντρου Απόφασης

- Η διαδικασία Decision Tree δημιουργεί ένα μοντέλο ταξινόμησης που βασίζεται σε δέντρο. Ταξινομεί τις περιπτώσεις σε ομάδες ή προβλέπει τιμές μιας εξαρτημένης (στόχου) μεταβλητής με βάση τις τιμές ανεξάρτητων (προγνωστικών) μεταβλητών. Η διαδικασία παρέχει εργαλεία επικύρωσης για διερευνητική και επιβεβαιωτική ανάλυση ταξινόμησης.

# Διαδικασία δημιουργίας δέντρου απόφασης

Η διαδικασία μπορεί να χρησιμοποιηθεί για:

- **Segmentation (Κατάτμηση):** Προσδιορίστε άτομα που είναι πιθανό να είναι μέλη μιας συγκεκριμένης ομάδας.
- **Stratification (Στρωμάτωση):** Καταχωρίστε τις περιπτώσεις σε μία από τις διάφορες κατηγορίες, όπως ομάδες υψηλού, μεσαίου και χαμηλού κινδύνου.
- **Prediction (Πρόβλεψη):** Δημιουργήστε κανόνες και χρησιμοποιήστε τους για να προβλέψετε μελλοντικά γεγονότα, όπως πχ: η πιθανότητα αθέτησης πληρωμών για ένα δάνειο ή η πιθανή αξία μεταπώλησης ενός οχήματος ή ενός σπιτιού.
- **Data reduction and variable screening (Μείωση δεδομένων και μεταβλητός έλεγχος):** Επιλέξτε ένα χρήσιμο υποσύνολο προγνωστικών από ένα μεγάλο σύνολο μεταβλητών για χρήση στην κατασκευή ενός επίσημου παραμετρικού μοντέλου.

# Διαδικασία δημιουργίας δέντρου απόφασης

- **Interaction identification (Ταυτοποίηση αλληλεπίδρασης):** Προσδιορίστε σχέσεις που αφορούν μόνο συγκεκριμένες υποομάδες και προσδιορίστε αυτές σε ένα επίσημο παραμετρικό μοντέλο.

- **Category merging and discretizing continuous variables (Συγχώνευση κατηγοριών και διακριτοποίηση συνεχών μεταβλητών):** Κωδικοποιήστε ξανά κατηγορίες προγνωστικών ομάδων και συνεχείς μεταβλητές με ελάχιστη απώλεια πληροφοριών.

# Παράδειγμα Σκεπτικού για υλοποίηση Δέντρου Απόφασης

- **Παράδειγμα:** Μια τράπεζα θέλει να κατηγοριοποιήσει τους αιτούντες πίστωση ανάλογα με το εάν αντιπροσωπεύουν ή όχι έναν εύλογο πιστωτικό κίνδυνο. Με βάση διάφορους παράγοντες, συμπεριλαμβανομένων των γνωστών αξιολογήσεων πιστοληπτικής ικανότητας παλαιότερων πελατών, μπορείτε να δημιουργήσετε ένα μοντέλο για να προβλέψετε εάν οι μελλοντικοί πελάτες είναι πιθανό να αθετήσουν τα δάνειά τους.

# Επεξεργασία πριν την υλοποίηση της επίλυσης…

- Μια ανάλυση που βασίζεται σε δέντρα παρέχει μερικά χαρακτηριστικά:

- Σας επιτρέπει να προσδιορίσετε ομοιογενείς ομάδες με υψηλό ή χαμηλό κίνδυνο.

- Διευκολύνει τη δημιουργία κανόνων για την πραγματοποίηση προβλέψεων για μεμονωμένες περιπτώσεις.

# Θεωρήσεις δεδομένων

- **Δεδομένα.** Οι εξαρτημένες και ανεξάρτητες μεταβλητές μπορεί να είναι:
- *Nominal (Ονομαστική):* Μια μεταβλητή μπορεί να αντιμετωπιστεί ως ονομαστική όταν οι τιμές της αντιπροσωπεύουν κατηγορίες χωρίς εγγενή κατάταξη (για παράδειγμα, το τμήμα της εταιρείας στην οποία εργάζεται ένας υπάλληλος). Παραδείγματα ονομαστικών μεταβλητών περιλαμβάνουν την περιοχή, τον ταχυδρομικό κώδικα και τη θρησκευτική πεποίθηση.
- *Ordinal:* Μια μεταβλητή μπορεί να αντιμετωπιστεί ως ordinal όταν οι τιμές της αντιπροσωπεύουν κατηγορίες με κάποια εγγενή κατάταξη (για παράδειγμα, επίπεδα ικανοποίησης από την υπηρεσία από πολύ δυσαρεστημένος σε πολύ ικανοποιημένος). Παραδείγματα τακτικών μεταβλητών περιλαμβάνουν βαθμολογίες στάσης που αντιπροσωπεύουν βαθμό ικανοποίησης ή αυτοπεποίθησης και βαθμολογίες βαθμολογίας προτίμησης.
- Scale: Μια μεταβλητή μπορεί να αντιμετωπιστεί ως scale (συνεχής) όταν οι τιμές της αντιπροσωπεύουν ταξινομημένες κατηγορίες με μια σημαντική μέτρηση, έτσι ώστε οι συγκρίσεις απόστασης μεταξύ των τιμών να είναι κατάλληλες. Παραδείγματα μεταβλητών κλίμακας περιλαμβάνουν την ηλικία σε χρόνια και το εισόδημα σε χιλιάδες δολάρια.

# Βάρη Συχνότητας, Υποθέσεις, Επίπεδο Μέτρησης

- **Βάρη συχνότητας (Frequency weights)**: Τα κλασματικά βάρη στρογγυλοποιούνται στον πλησιέστερο ακέραιο. Έτσι, σε περιπτώσεις με τιμή βάρους μικρότερη από 0,5 αποδίδεται βάρος 0 και επομένως εξαιρούνται από την ανάλυση.
- **Υποθέσεις (Assumptions):** Αυτή η διαδικασία προϋποθέτει ότι το κατάλληλο επίπεδο μέτρησης έχει εκχωρηθεί σε όλες τις μεταβλητές ανάλυσης και ορισμένα χαρακτηριστικά υποθέτουν ότι όλες οι τιμές της εξαρτημένης μεταβλητής που περιλαμβάνονται στην ανάλυση έχουν καθορισμένες ετικέτες τιμών.
- **Επίπεδο μέτρησης (Measurement level):** Το επίπεδο μέτρησης επηρεάζει τους υπολογισμούς του δέντρου. Έτσι, σε όλες τις μεταβλητές θα πρέπει να εκχωρηθεί το κατάλληλο επίπεδο μέτρησης. Από προεπιλογή, οι αριθμητικές μεταβλητές υποτίθεται ότι είναι scale και οι μεταβλητές συμβολοσειράς θεωρούνται nominal οι οποίες μπορεί να μην αντικατοπτρίζουν με ακρίβεια το πραγματικό επίπεδο μέτρησης. Ένα εικονίδιο δίπλα σε κάθε μεταβλητή στη λίστα μεταβλητών προσδιορίζει τον τύπο της μεταβλητής.

| Icon | Measurement level |
|---|---|
|  | Scale |
|  | Nominal |
|  | Ordinal |

# Value Labels

- **Ετικέτες αξίας (Value Labels):** Η διεπαφή του πλαισίου διαλόγου για αυτήν τη διαδικασία υποθέτει ότι είτε όλες οι τιμές που δεν λείπουν μιας κατηγορικής (nominal, ordinal) εξαρτημένης μεταβλητής έχουν καθορισμένες ετικέτες τιμών ή καμία από αυτές δεν έχει.

- Μπορείτε να χρησιμοποιήσετε το **Define Variable Properties** για να σας βοηθήσει στη διαδικασία καθορισμού τόσο των ετικετών επιπέδου μέτρησης όσο και τιμών.

# Βήματα για δημιουργία Δέντρου Απόφασης

- Από τα μενού επιλέξτε:

  Analyze→ Classify→ Tree

1. Επιλέξτε μια dependent (εξαρτημένη) μεταβλητή.
2. Επιλέξτε μία ή περισσότερες independent (ανεξάρτητες) μεταβλητές.
3. Επιλέξτε growing method.

# Growing Methods

- ***CHAID**:* Αυτόματη ανίχνευση αλληλεπίδρασης chi-squared. Σε κάθε βήμα, το CHAID επιλέγει την ανεξάρτητη (προβλεπόμενη) μεταβλητή που έχει την ισχυρότερη αλληλεπίδραση με την εξαρτημένη μεταβλητή. Οι κατηγορίες κάθε προγνωστικού παράγοντα συγχωνεύονται εάν δεν διαφέρουν σημαντικά σε σχέση με την εξαρτημένη μεταβλητή.
- ***CRT**:* Δέντρα ταξινόμησης και παλινδρόμησης. Το CRT διαχωρίζει τα δεδομένα σε τμήματα που είναι όσο το δυνατόν πιο ομοιογενή σε σχέση με την εξαρτημένη μεταβλητή. Ένας τερματικός κόμβος στον οποίο όλες οι περιπτώσεις έχουν την ίδια τιμή για την εξαρτημένη μεταβλητή είναι ένας ομοιογενής, "καθαρός" κόμβος.
- ***QUEST***: Γρήγορο, αμερόληπτο, αποτελεσματικό στατιστικό δέντρο. Μια μέθοδος που είναι γρήγορη και αποφεύγει την προκατάληψη άλλων μεθόδων υπέρ προγνωστικών με πολλές κατηγορίες. Το QUEST μπορεί να καθοριστεί μόνο εάν η εξαρτημένη μεταβλητή είναι nominal.

# Χαρακτηριστικά Growing Methods for Decision Tree

| Feature | CHAID* | CRT | QUEST |
|---|---|---|---|
| Chi-square-based** | X | | |
| Surrogate independent (predictor) variables | | X | X |
| Tree pruning | | X | X |
| Multiway node splitting | X | | |
| Binary node splitting | | X | X |
| Influence variables | X | X | |
| Prior probabilities | | X | X |
| Misclassification costs | X | X | X |
| Fast calculation | X | | X |

*Includes Exhaustive CHAID.

**QUEST also uses a chi-square measure for nominal independent variables.

# Αναλυτικά για τα Δέντρα Απόφασης

# Επιλογή Κατηγοριών για τις μεταβλητές

- Από το μενού επιλέξτε:

1. Analyze → Classify → Tree
2. Στο κύριο πλαίσιο διαλόγου Δέντρο αποφάσεων, επιλέξτε μια κατηγορική (nominal, ordinal) εξαρτημένη μεταβλητή με δύο ή περισσότερες ετικέτες καθορισμένων τιμών.
3. Κάντε κλικ στο Categories.

# Επικύρωση Validation

- Το validation σας επιτρέπει να αξιολογήσετε πόσο καλά η δομή των δέντρων σας γενικεύεται σε μεγαλύτερο πληθυσμό. Δύο μέθοδοι επικύρωσης είναι διαθέσιμες: διασταυρούμενη επικύρωση cross validation και επικύρωση διαχωρισμού δειγμάτων split-sample validation.
- ***Cross validation***: διαιρεί το δείγμα σε έναν αριθμό υποδειγμάτων. Στη συνέχεια δημιουργούνται μοντέλα δέντρων. Παράγει ένα ενιαίο, τελικό μοντέλο δέντρου. Η διασταυρούμενη εκτίμηση κινδύνου για το τελικό δέντρο υπολογίζεται ως ο μέσος όρος των κινδύνων για όλα τα δέντρα.
- **Split-Sample Validation**: το μοντέλο δημιουργείται χρησιμοποιώντας ένα δείγμα εκπαίδευσης και δοκιμάζεται σε ένα δείγμα διατήρησης. θα πρέπει να χρησιμοποιείται με προσοχή σε μικρά αρχεία δεδομένων (αρχεία δεδομένων με μικρό αριθμό περιπτώσεων). Τα μικρά μεγέθη δειγμάτων εκπαίδευσης μπορεί να αποδώσουν φτωχά μοντέλα, καθώς μπορεί να μην υπάρχουν αρκετές περιπτώσεις σε ορισμένες κατηγορίες για να αναπτυχθεί επαρκώς το δέντρο.

Από το μενού επιλέξτε:
1. Analyze→ Classify → Tree
2. Κάντε κλικ στο validation
3. Επιλέξτε Cross validation ή Split-sample Validation

# Επιλογή κριτηρίου

- Για να καθορίσετε το κριτήριο **CHAID**:

1. Analyze → Classify → Tree
2. Στο κύριο παράθυρο διαλόγου Decision Tree, επιλέξτε CHAID
3. Κάντε κλικ στο Criteria
4. Κάντε κλικ στο CHAID

- Για να καθορίσετε το κριτήριο **CRT**:

1. Analyze → Classify → Tree
2. Στο κύριο παράθυρο διαλόγου Decision Tree, επιλέξτε CRT
3. Κάντε κλικ στο Criteria
4. Κάντε κλικ στο CRT

- Για να καθορίσετε το κριτήριο **QUEST**:

1. Analyze → Classify → Tree
2. Στο κύριο παράθυρο διαλόγου Decision Tree, επιλέξτε QUEST
3. Κάντε κλικ στο Criteria
4. Κάντε κλικ στο QUEST

# Κλάδεμα Δέντρων Pruning Trees

- Με τις μεθόδους CRT και QUEST, μπορείτε να αποφύγετε την υπερβολική προσαρμογή του μοντέλου κλαδεύοντας το δέντρο: το δέντρο αναπτύσσεται μέχρι να ικανοποιηθούν τα κριτήρια διακοπής και, στη συνέχεια, κόβεται αυτόματα στο μικρότερο υποδέντρο με βάση την καθορισμένη μέγιστη διαφορά κινδύνου. Η τιμή κινδύνου εκφράζεται σε τυπικά σφάλματα. Η προεπιλογή είναι 1. Η τιμή πρέπει να είναι μη αρνητική. Για να αποκτήσετε το υποδέντρο με τον ελάχιστο κίνδυνο, καθορίστε το 0.

- Για να κλαδέψετε ένα δέντρο:

1. Analyze → Classify → Tree
2. Στο κύριο πλαίσιο διαλόγου Decision Tree, για τη μέθοδο ανάπτυξης, επιλέξτε CRT ή QUEST.
3. Κάντε κλικ στην επιλογή Criteria.
4. Κάντε κλικ στην καρτέλα Pruning.

# Tree Display

- **Orientation (Προσανατολισμός):** Το δέντρο μπορεί να εμφανίζεται από πάνω προς τα κάτω με τον κόμβο ρίζας στην κορυφή, από αριστερά προς τα δεξιά ή από δεξιά προς τα αριστερά.
- **Node Contents (Περιεχόμενα κόμβου):** Οι κόμβοι μπορούν να εμφανίζουν πίνακες, γραφήματα ή και τα δύο. Για τις κατηγορικές εξαρτημένες μεταβλητές, οι πίνακες εμφανίζουν μετρήσεις συχνότητας και ποσοστά και τα γραφήματα είναι γραφήματα ράβδων. Για μεταβλητές που εξαρτώνται από κλίμακα, οι πίνακες εμφανίζουν τα μέσα, τις τυπικές αποκλίσεις, τον αριθμό των περιπτώσεων και τις προβλεπόμενες τιμές και τα γραφήματα είναι ιστογράμματα.
- **Scale (Κλίμακα):** Από προεπιλογή, τα μεγάλα δέντρα μειώνονται αυτόματα σε μια προσπάθεια να χωρέσουν το δέντρο στη σελίδα. Μπορείτε να καθορίσετε ένα ποσοστό προσαρμοσμένης κλίμακας έως και 200%.
- **Tree in Table Format (Δέντρο σε μορφή πίνακα):** Συνοπτικές πληροφορίες για κάθε κόμβο στο δέντρο, συμπεριλαμβανομένου του αριθμού γονικού κόμβου, των στατιστικών ανεξάρτητων μεταβλητών, της τιμής ανεξάρτητης μεταβλητής για τον κόμβο, της μέσης και τυπικής απόκλισης για τις μεταβλητές που εξαρτώνται από την κλίμακα ή των μετρήσεων και των ποσοστών για τις κατηγορικές εξαρτημένες μεταβλητές.

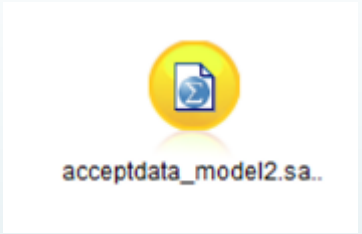
Από το μενού επιλέξτε:

1. Analyze → Classify → Tree
2. Στο κύριο παράθυρο διαλόγου Δέντρο αποφάσεων, κάντε κλικ στην επιλογή Output
3. Κάντε κλικ στην καρτέλα Tree.

# Ενδεικτικό Παράδειγμα

# Δημιουργία μοντέλου

Ξεκινώντας, θα επιλέξω και θα εισάγω το αρχείο .sav στο SPSS του υπολογιστή μου.




Έπειτα, παρουσιάζεται το dataset στο SPSS

Table (26 fields, 5,056 records)

File Edit Generate

Table Annotations

| | Admission | alumni_mtg | facebook_page | overnight | fin_aid | info_packet_mail | info_packet_email | enroll |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.700 | N | Y | N | 1% - 25% | N | Y | N |
| 2 | 0.897 | N | Y | N | 0% | N | Y | N |
| 3 | 0.394 | Y | Y | Y | 0% | Y | Y | Y |
| 4 | 0.394 | Y | Y | N | 0% | Y | Y | N |
| 5 | 0.394 | Y | Y | N | 0% | Y | Y | Y |
| 6 | 0.897 | Y | Y | Y | 0% | Y | Y | Y |
| 7 | 0.884 | N | Y | Y | 0% | Y | Y | N |
| 8 | 0.884 | N | Y | N | 0% | Y | Y | N |
| 9 | 0.897 | N | Y | N | 0% | N | Y | N |
| 10 | 0.454 | N | Y | N | 0% | Y | Y | N |

# Δημιουργία μοντέλου

- Το επόμενο βήμα στη διαδικασία είναι η ανάγνωση των δεδομένων καθορίζοντας τις ιδιότητες δεδομένων για κάθε πεδίο.
- Δηλαδή αν θα είναι scale, nominal, ordinal καθώς επίσης και αν θα έχουν value labels.

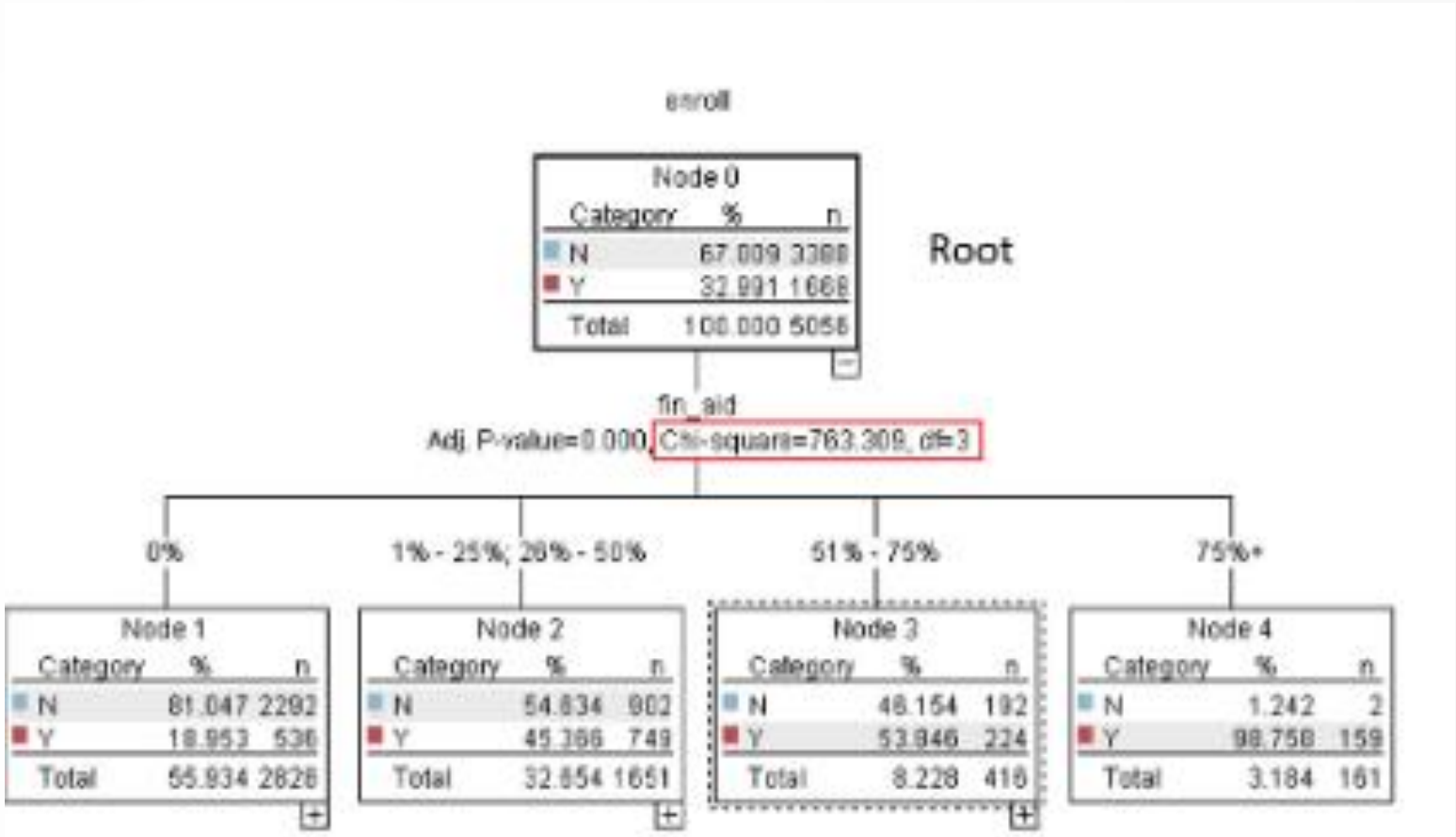
# Δημιουργία μοντέλου

- Έπειτα, επιλέγουμε κριτήριο με το οποίο θα δημιουργήσουμε το decision Tree μας.

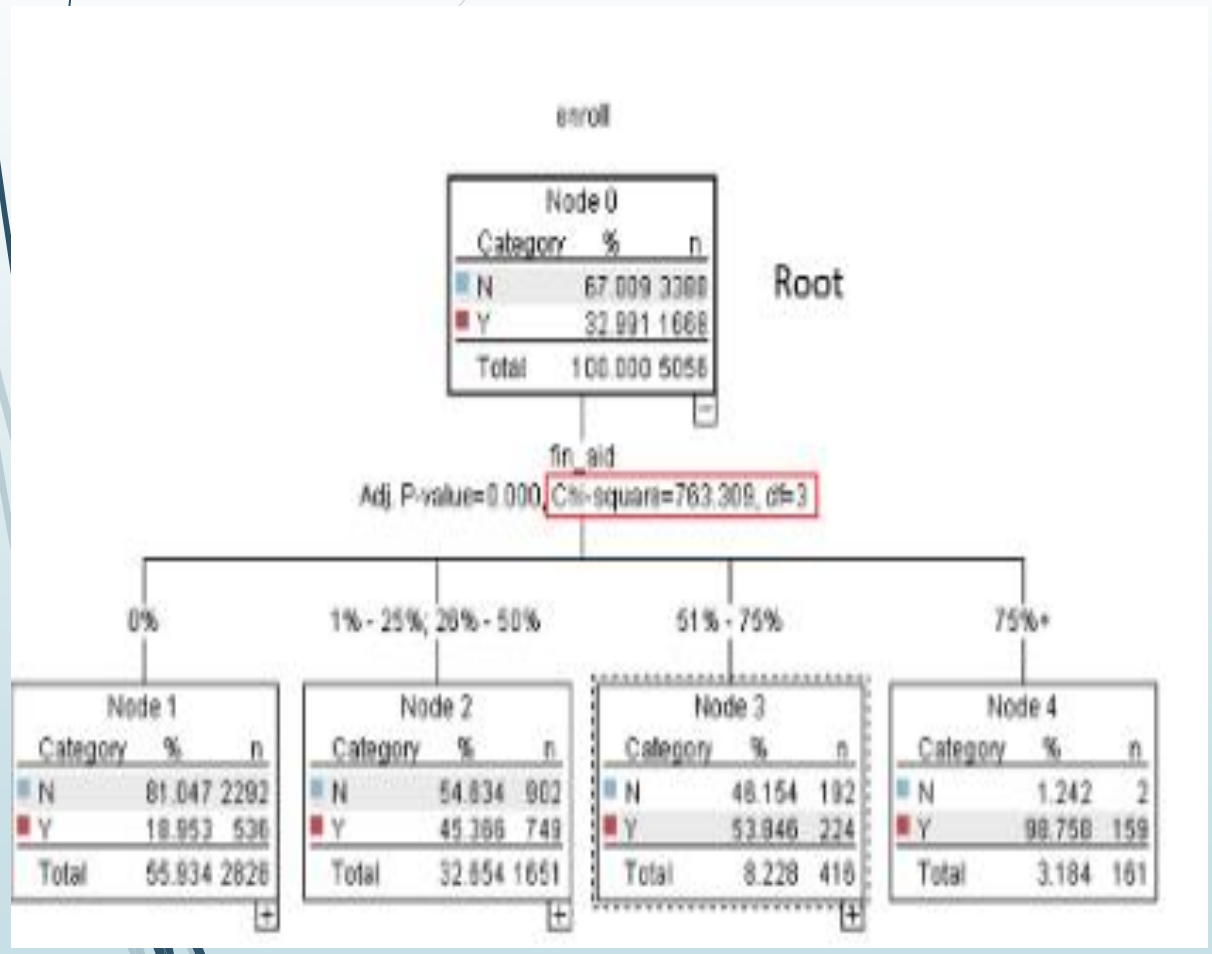
1. Analyze → Classify → Tree
2. Στο κύριο παράθυρο διαλόγου Decision Tree, επιλέξτε το επιθυμητό κριτήριο (πχ. CHAID)
3. Κάνουμε παραμετροποιήσεις όσο αφορά validation, criteria, output.

# Ερμηνεία των αποτελεσμάτων Ι

- Το δέντρο απόφασης ξεκινά με τον ριζικό κόμβο, ο οποίος απλώς δείχνει την κατανομή του πεδίου αποτελέσματος, το οποίο όπως γνωρίζουμε είναι εγγραφή. Στη συνέχεια, τα δεδομένα διαχωρίζονται με βάση τη στατιστική σημασία από τον προγνωστικό παράγοντα με την ισχυρότερη σχέση με το πεδίο-στόχο, την οικονομική βοήθεια σε αυτήν την περίπτωση. Και μπορείτε να δείτε ότι υπάρχουν πέντε «κουβάδες» στους οποίους έχει χωριστεί η οικονομική βοήθεια (0%, 1%-25%, 26%-5-%, 51%-75% και 75% +). Εξετάζοντας εκείνους τους φοιτητές στους οποίους προσφέρθηκε ένα πακέτο οικονομικής βοήθειας 51%-75%, το μοντέλο ήταν σε θέση να προβλέψει ότι αυτοί οι φοιτητές θα εγγραφούν περίπου στο 54% του χρόνου. Αυτή η πρόβλεψη εφαρμόστηκε σε 416 μαθητές και το μοντέλο ήταν ακριβές 224 φορές.



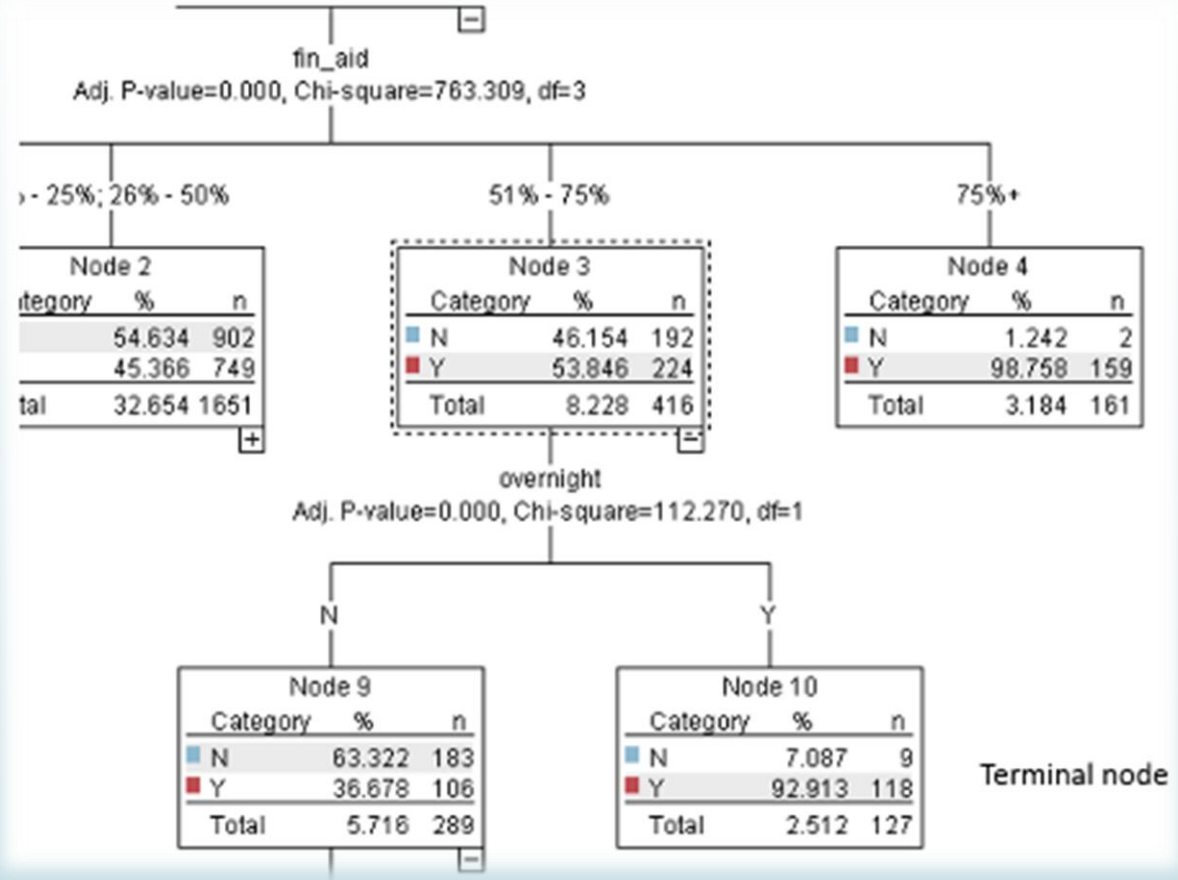
# Ερμηνεία των αποτελεσμάτων ΙΙ




- Καθώς συνεχίζουμε να κατεβαίνουμε το δέντρο, βλέπουμε ότι η επόμενη πιο σημαντική μεταβλητή είναι μια ολονύκτια επίσκεψη. Εάν προσφερόταν σε έναν μαθητή ένα πακέτο οικονομικής βοήθειας 51%-75% και πραγματοποιούσε επίσης μια ολονύκτια επίσκεψη, μπορούσαμε να προβλέψουμε με ακρίβεια ότι θα εγγραφόταν περίπου στο 93% του χρόνου. Εναλλακτικά, εάν οι μαθητές δεν πραγματοποιούσαν μια ολονύκτια επίσκεψη, προβλέψαμε ότι δεν θα εγγραφούν στο 63% του χρόνου. Αυτός ο κανόνας ίσχυε για 289 μαθητές και ήμασταν ακριβείς περίπου 183 φορές.
- Και ακριβώς έτσι συνεχίζουμε να κατεβαίνουμε το δέντρο προς την επόμενη πιο σημαντική μεταβλητή μέχρι να φτάσουμε σε έναν τερματικό κόμβο, που σημαίνει ότι η πρόβλεψη έχει τελειώσει

# Ερμηνεία των αποτελεσμάτων III

- Αυτό ήταν ένα απλό δέντρο αποφάσεων με στόχο να δείξει ποιες μεταβλητές μας βοηθούν να προβλέψουμε με ακρίβεια την εγγραφή των μαθητών. Λάβετε υπόψη ότι η προγνωστική ανάλυση μπορεί να εφαρμοστεί σε διάφορους κλάδους, όπως η εκπαίδευση, το λιανικό εμπόριο, η υγειονομική περίθαλψη και η χρηματοδότηση για να αναφέρουμε μόνο μερικές.



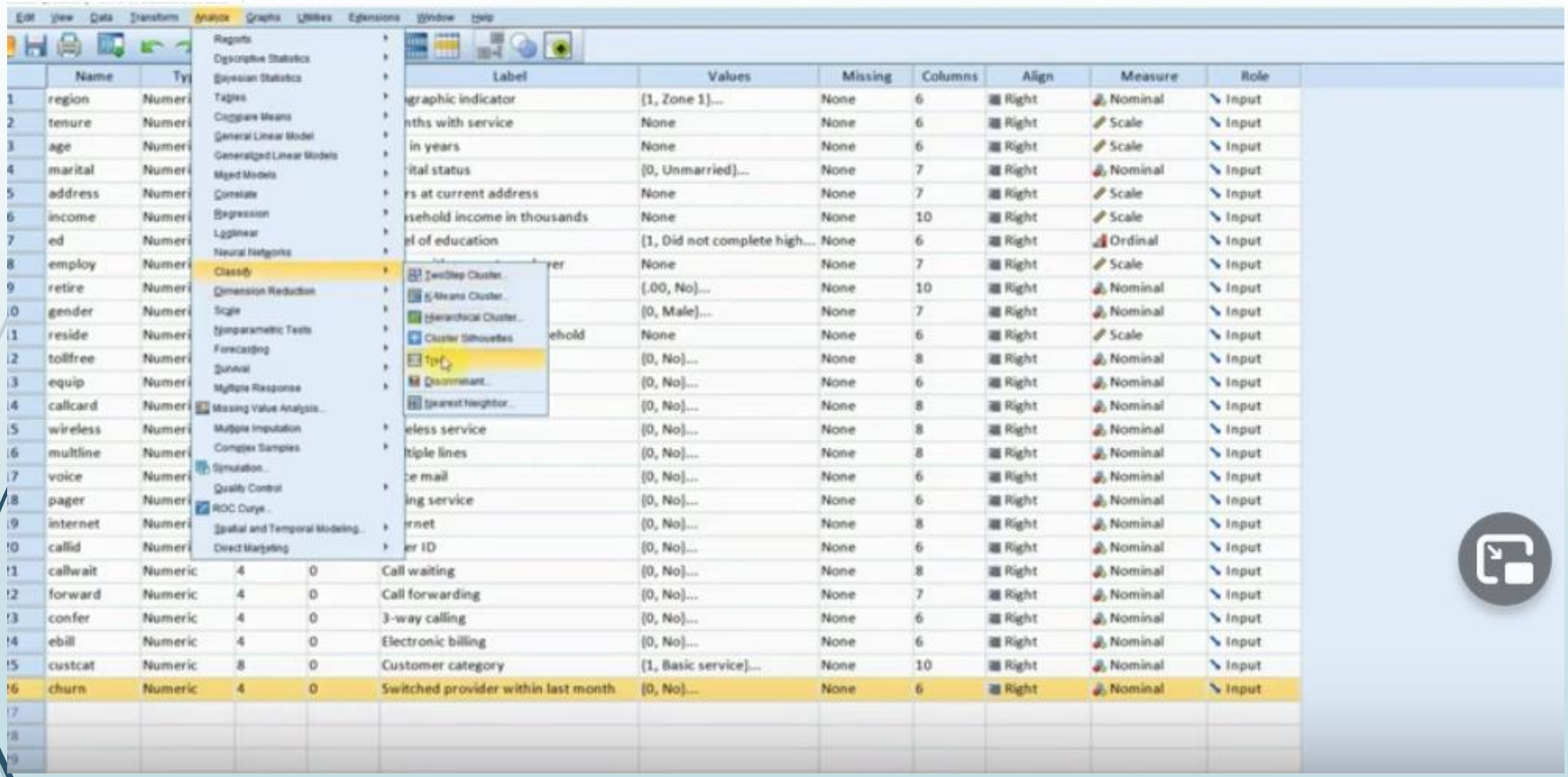
# Μηχανική μάθηση με το SPSS

Αυτόματη Κατηγοριοποίηση

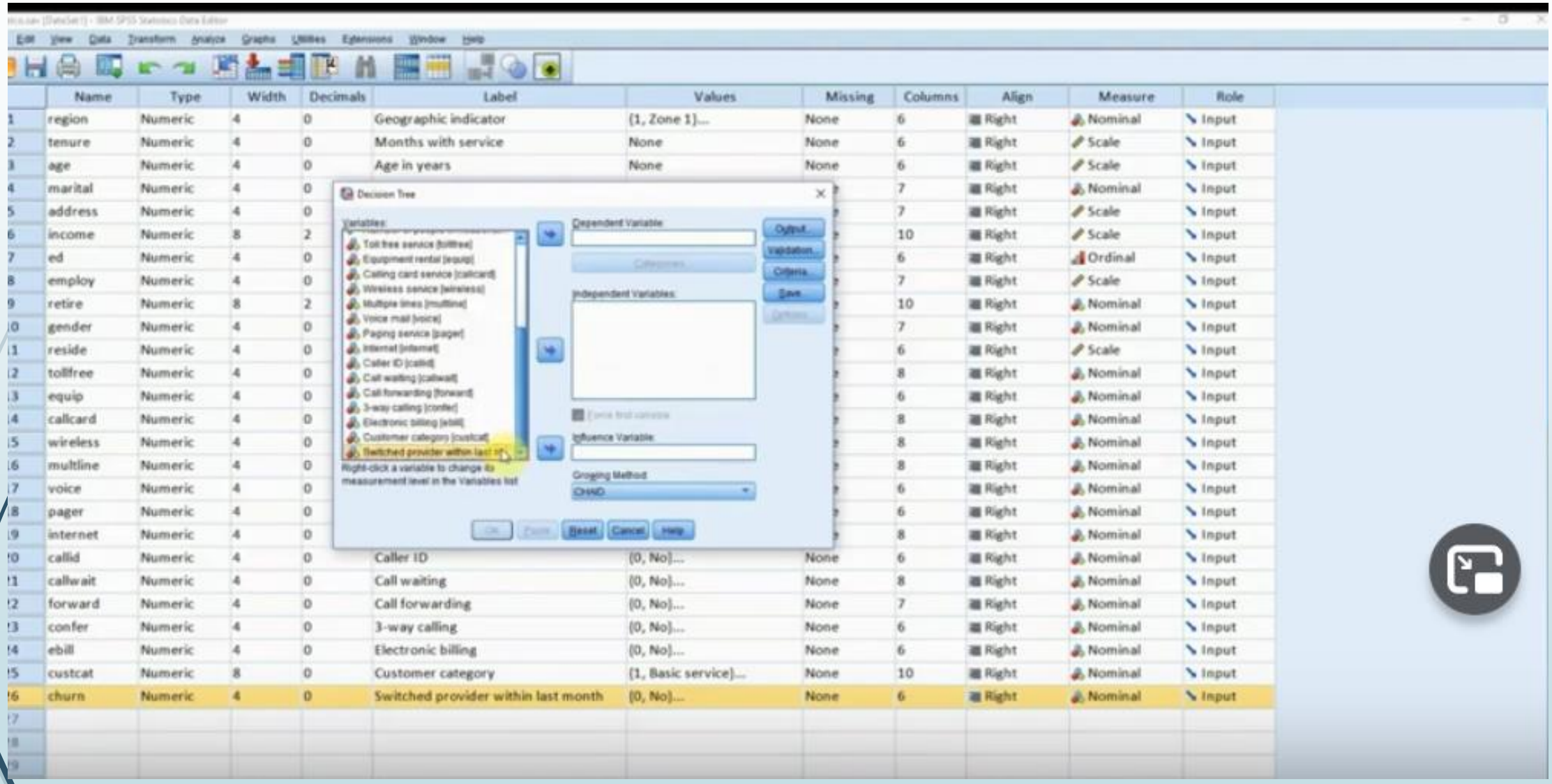
Δεδομένων

Με στόχο την επιλογή παραμέτρων για την πρόβλεψη ενός αποτελέσματος

# Τα δεδομένα μας…

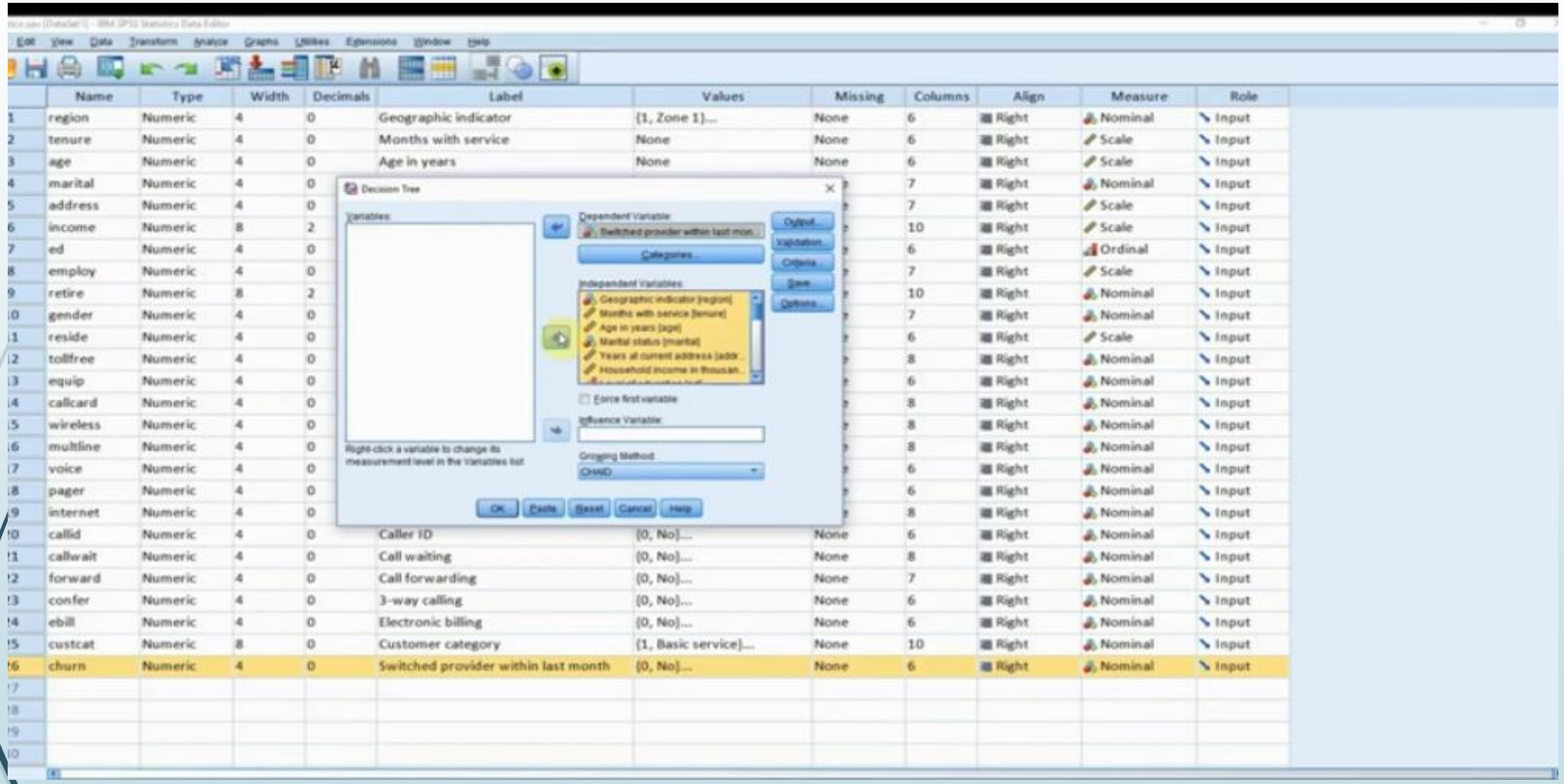


# Η εξαρτημένη μεταβλητή / στόχος



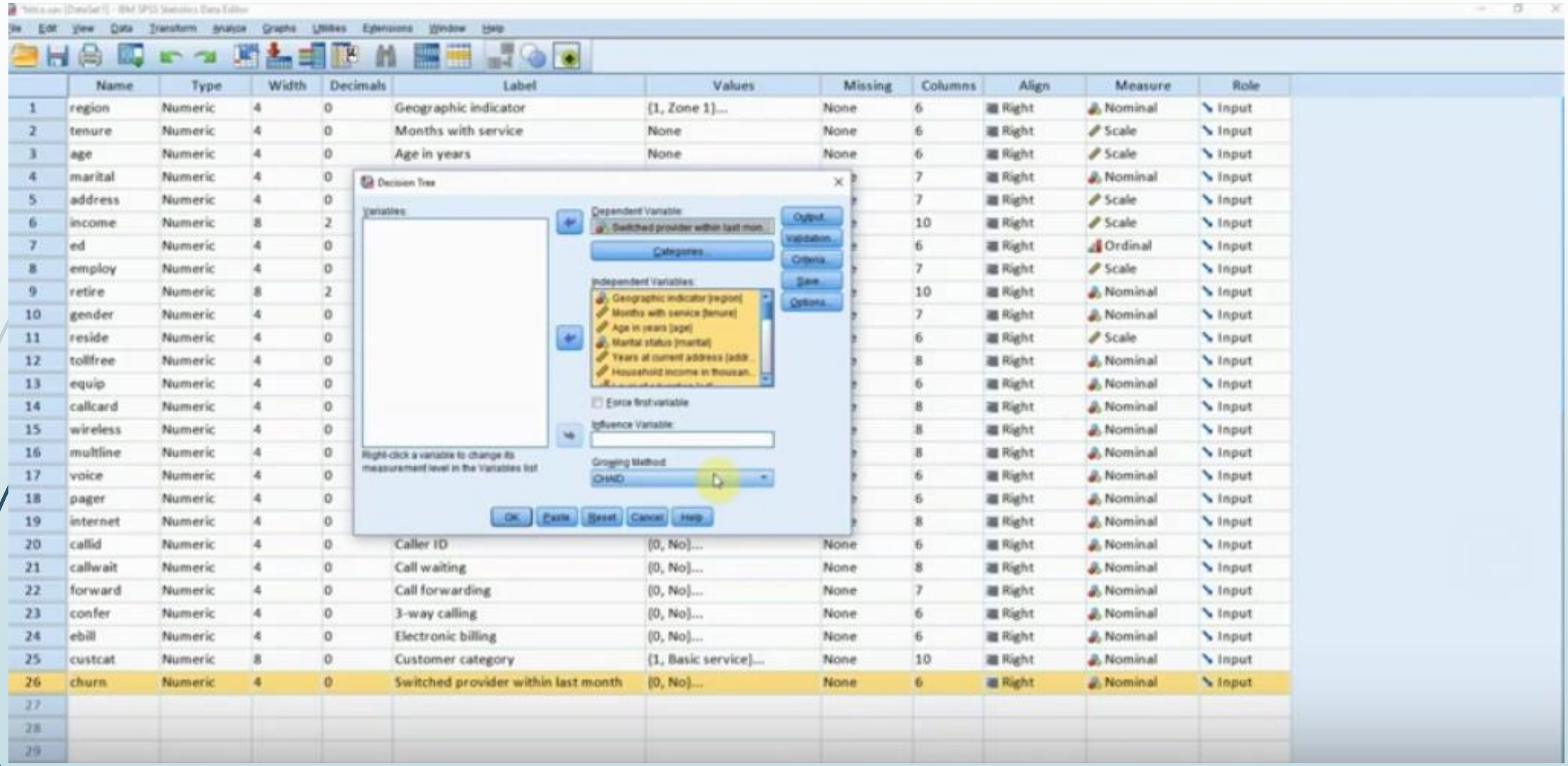
| | Name | Type | Width | Decimals | Label | Values | Missing | Columns | Align | Measure | Role |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | region | Numeric | 4 | 0 | Geographic indicator | {1, Zone 1}... | None | 6 | Right | Nominal | Input |
| 2 | tenure | Numeric | 4 | 0 | Months with service | None | None | 6 | Right | Scale | Input |
| 3 | age | Numeric | 4 | 0 | Age in years | None | None | 6 | Right | Scale | Input |
| 4 | marital | Numeric | 4 | 0 | | | | 7 | Right | Nominal | Input |
| 5 | address | Numeric | 4 | 0 | | | | 7 | Right | Scale | Input |
| 6 | income | Numeric | 8 | 2 | | | | 10 | Right | Scale | Input |
| 7 | ed | Numeric | 4 | 0 | | | | 6 | Right | Ordinal | Input |
| 8 | employ | Numeric | 4 | 0 | | | | 7 | Right | Scale | Input |
| 9 | retire | Numeric | 8 | 2 | | | | 10 | Right | Nominal | Input |
| 10 | gender | Numeric | 4 | 0 | | | | 7 | Right | Nominal | Input |
| 11 | reside | Numeric | 4 | 0 | | | | 6 | Right | Scale | Input |
| 12 | tollfree | Numeric | 4 | 0 | | | | 8 | Right | Nominal | Input |
| 13 | equip | Numeric | 4 | 0 | | | | 6 | Right | Nominal | Input |
| 14 | callcard | Numeric | 4 | 0 | | | | 8 | Right | Nominal | Input |
| 15 | wireless | Numeric | 4 | 0 | | | | 8 | Right | Nominal | Input |
| 16 | multline | Numeric | 4 | 0 | | | | 8 | Right | Nominal | Input |
| 17 | voice | Numeric | 4 | 0 | | | | 6 | Right | Nominal | Input |
| 18 | pager | Numeric | 4 | 0 | | | | 6 | Right | Nominal | Input |
| 19 | internet | Numeric | 4 | 0 | | | | 8 | Right | Nominal | Input |
| 20 | callid | Numeric | 4 | 0 | Caller ID | {0, No}... | None | 6 | Right | Nominal | Input |
| 21 | callwait | Numeric | 4 | 0 | Call waiting | {0, No}... | None | 8 | Right | Nominal | Input |
| 22 | forward | Numeric | 4 | 0 | Call forwarding | {0, No}... | None | 7 | Right | Nominal | Input |
| 23 | confer | Numeric | 4 | 0 | 3-way calling | {0, No}... | None | 6 | Right | Nominal | Input |
| 24 | ebill | Numeric | 4 | 0 | Electronic billing | {0, No}... | None | 6 | Right | Nominal | Input |
| 25 | custcat | Numeric | 8 | 0 | Customer category | {1, Basic service}... | None | 10 | Right | Nominal | Input |
| 26 | churn | Numeric | 4 | 0 | Switched provider within last month | {0, No}... | None | 6 | Right | Nominal | Input |

# Συσχέτιση με ΟΛΕΣ τις ανεξάρτητες…

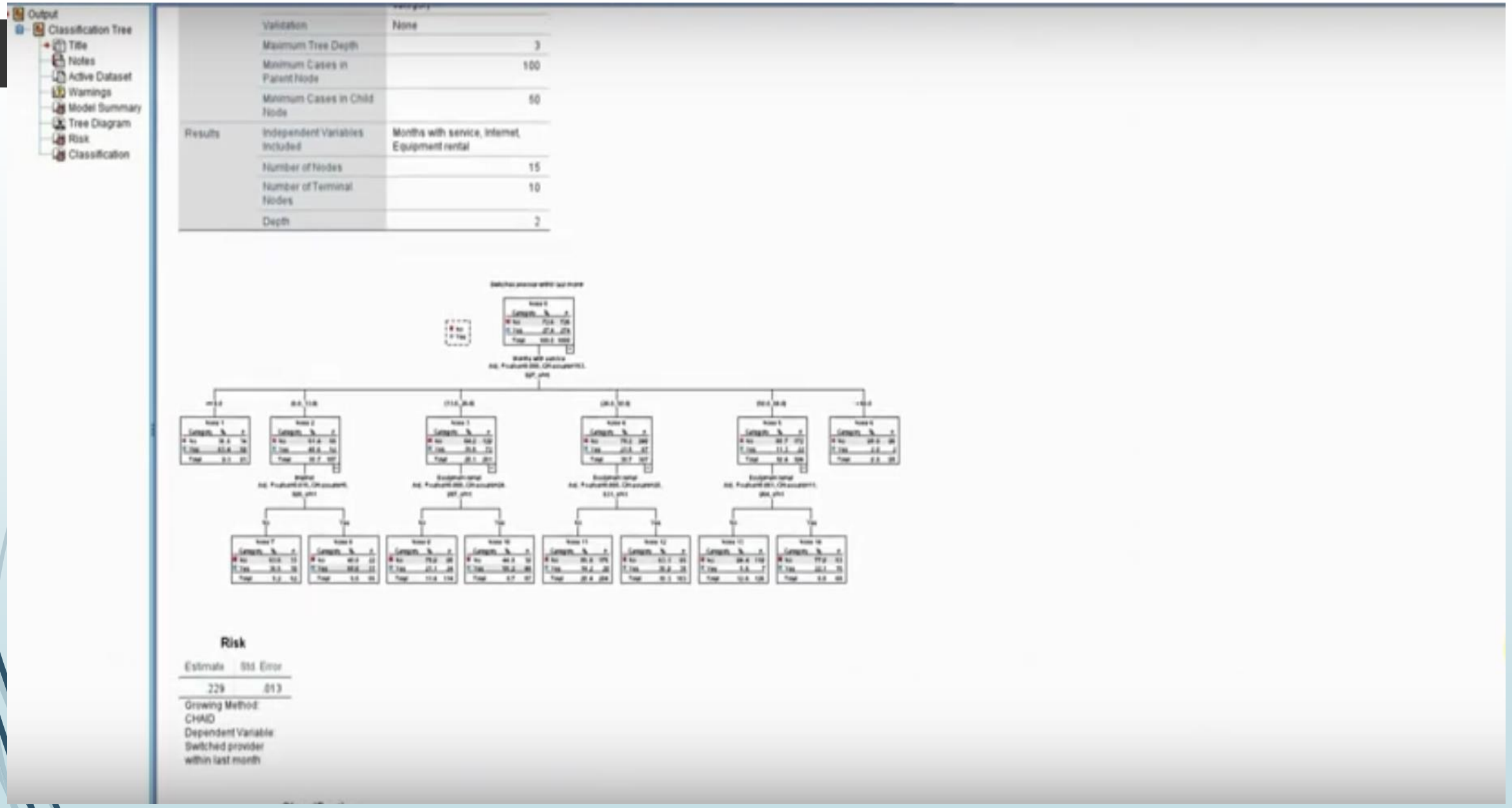


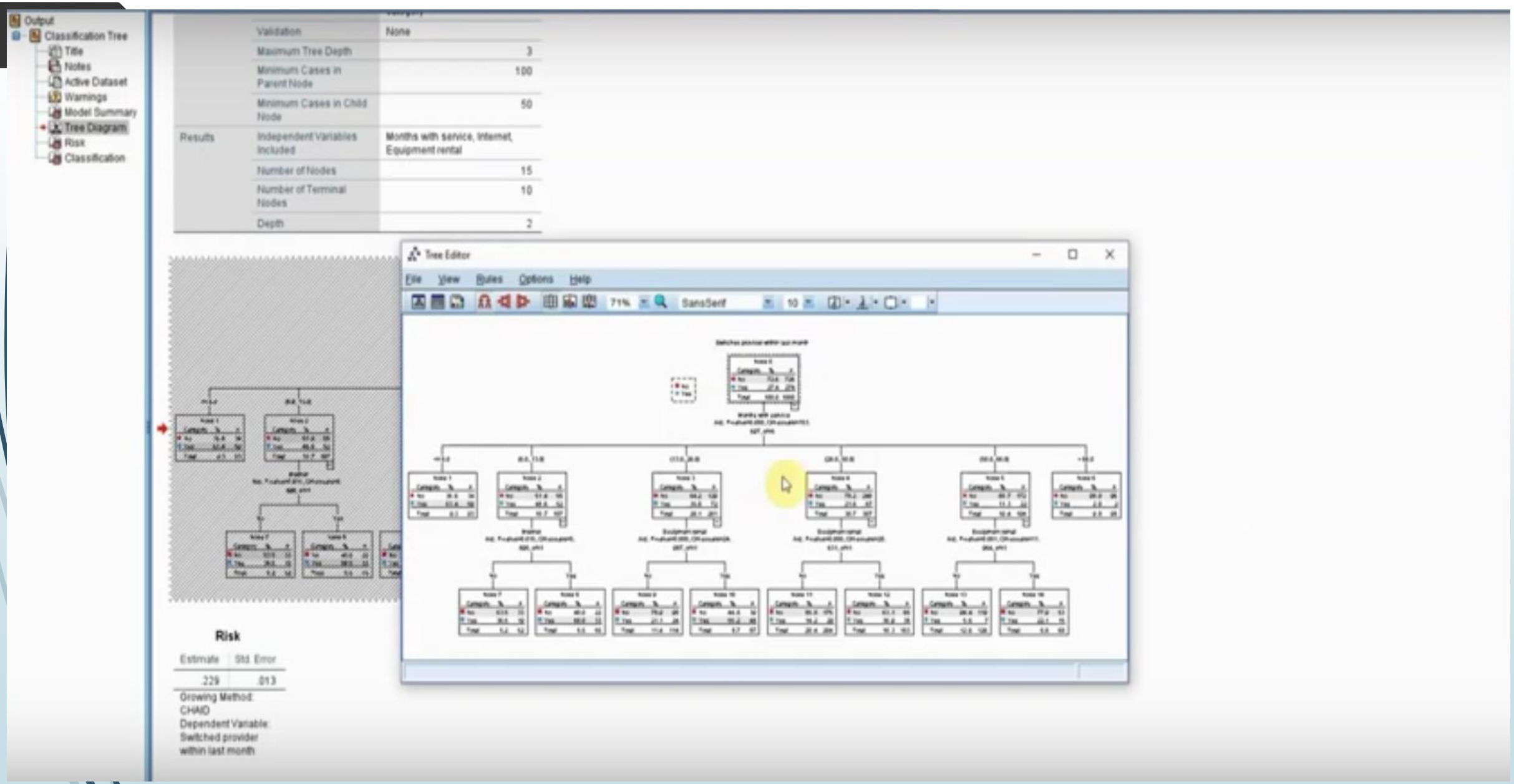
| | Name | Type | Width | Decimals | Label | Values | Missing | Columns | Align | Measure | Role |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | region | Numeric | 4 | 0 | Geographic indicator | {1, Zone 1}... | None | 6 | Right | Nominal | Input |
| 2 | tenure | Numeric | 4 | 0 | Months with service | None | None | 6 | Right | Scale | Input |
| 3 | age | Numeric | 4 | 0 | Age in years | None | None | 6 | Right | Scale | Input |
| 4 | marital | Numeric | 4 | 0 | | | | 7 | Right | Nominal | Input |
| 5 | address | Numeric | 4 | 0 | | | | 7 | Right | Scale | Input |
| 6 | income | Numeric | 8 | 2 | | | | 10 | Right | Scale | Input |
| 7 | ed | Numeric | 4 | 0 | | | | 6 | Right | Ordinal | Input |
| 8 | employ | Numeric | 4 | 0 | | | | 7 | Right | Scale | Input |
| 9 | retire | Numeric | 8 | 2 | | | | 10 | Right | Nominal | Input |
| 10 | gender | Numeric | 4 | 0 | | | | 7 | Right | Nominal | Input |
| 11 | reside | Numeric | 4 | 0 | | | | 6 | Right | Scale | Input |
| 12 | tollfree | Numeric | 4 | 0 | | | | 8 | Right | Nominal | Input |
| 13 | equip | Numeric | 4 | 0 | | | | 6 | Right | Nominal | Input |
| 14 | callcard | Numeric | 4 | 0 | | | | 8 | Right | Nominal | Input |
| 15 | wireless | Numeric | 4 | 0 | | | | 8 | Right | Nominal | Input |
| 16 | multline | Numeric | 4 | 0 | | | | 8 | Right | Nominal | Input |
| 17 | voice | Numeric | 4 | 0 | | | | 6 | Right | Nominal | Input |
| 18 | pager | Numeric | 4 | 0 | | | | 6 | Right | Nominal | Input |
| 19 | internet | Numeric | 4 | 0 | | | | 8 | Right | Nominal | Input |
| 20 | callid | Numeric | 4 | 0 | Caller ID | {0, No}... | None | 6 | Right | Nominal | Input |
| 21 | callwait | Numeric | 4 | 0 | Call waiting | {0, No}... | None | 8 | Right | Nominal | Input |
| 22 | forward | Numeric | 4 | 0 | Call forwarding | {0, No}... | None | 7 | Right | Nominal | Input |
| 23 | confer | Numeric | 4 | 0 | 3-way calling | {0, No}... | None | 6 | Right | Nominal | Input |
| 24 | ebill | Numeric | 4 | 0 | Electronic billing | {0, No}... | None | 6 | Right | Nominal | Input |
| 25 | custcat | Numeric | 8 | 0 | Customer category | {1, Basic service}... | None | 10 | Right | Nominal | Input |
| 26 | churn | Numeric | 4 | 0 | Switched provider within last month | {0, No}... | None | 6 | Right | Nominal | Input |

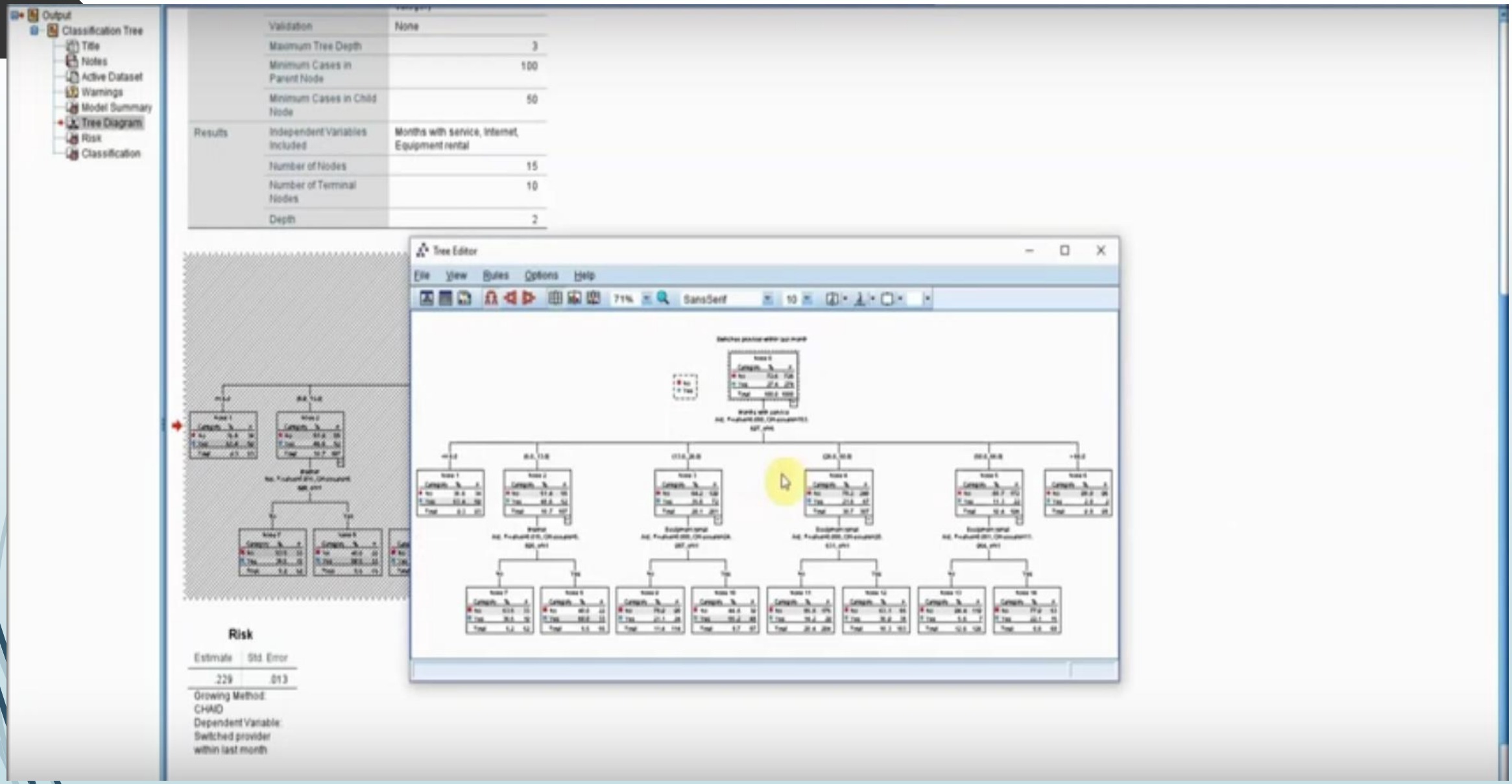
# Επιλογή Αλγορίθμου…

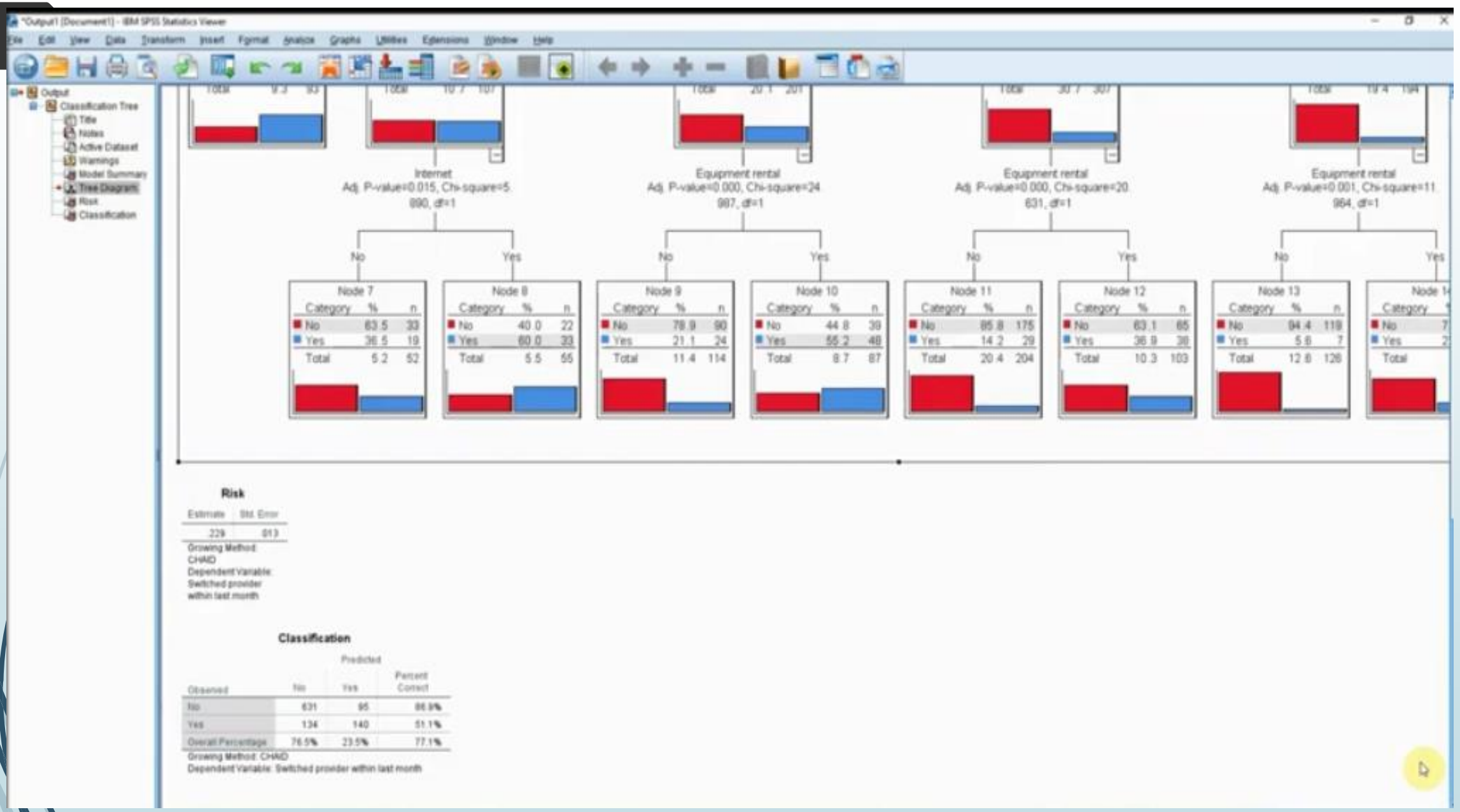


| | Name | Type | Width | Decimals | Label | Values | Missing | Columns | Align | Measure | Role |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | region | Numeric | 4 | 0 | Geographic indicator | {1, Zone 1}... | None | 6 | Right | Nominal | Input |
| 2 | tenure | Numeric | 4 | 0 | Months with service | None | None | 6 | Right | Scale | Input |
| 3 | age | Numeric | 4 | 0 | Age in years | None | None | 6 | Right | Scale | Input |
| 4 | marital | Numeric | 4 | 0 | | | | 7 | Right | Nominal | Input |
| 5 | address | Numeric | 4 | 0 | | | | 7 | Right | Scale | Input |
| 6 | income | Numeric | 8 | 2 | | | | 10 | Right | Scale | Input |
| 7 | ed | Numeric | 4 | 0 | | | | 6 | Right | Ordinal | Input |
| 8 | employ | Numeric | 4 | 0 | | | | 7 | Right | Scale | Input |
| 9 | retire | Numeric | 8 | 2 | | | | 10 | Right | Nominal | Input |
| 10 | gender | Numeric | 4 | 0 | | | | 7 | Right | Nominal | Input |
| 11 | reside | Numeric | 4 | 0 | | | | 6 | Right | Scale | Input |
| 12 | tollfree | Numeric | 4 | 0 | | | | 8 | Right | Nominal | Input |
| 13 | equip | Numeric | 4 | 0 | | | | 6 | Right | Nominal | Input |
| 14 | callcard | Numeric | 4 | 0 | | | | 8 | Right | Nominal | Input |
| 15 | wireless | Numeric | 4 | 0 | | | | 8 | Right | Nominal | Input |
| 16 | multline | Numeric | 4 | 0 | | | | 8 | Right | Nominal | Input |
| 17 | voice | Numeric | 4 | 0 | | | | 6 | Right | Nominal | Input |
| 18 | pager | Numeric | 4 | 0 | | | | 6 | Right | Nominal | Input |
| 19 | internet | Numeric | 4 | 0 | | | | 8 | Right | Nominal | Input |
| 20 | callid | Numeric | 4 | 0 | Caller ID | {0, No}... | None | 6 | Right | Nominal | Input |
| 21 | callwait | Numeric | 4 | 0 | Call waiting | {0, No}... | None | 8 | Right | Nominal | Input |
| 22 | forward | Numeric | 4 | 0 | Call forwarding | {0, No}... | None | 7 | Right | Nominal | Input |
| 23 | confer | Numeric | 4 | 0 | 3-way calling | {0, No}... | None | 6 | Right | Nominal | Input |
| 24 | ebill | Numeric | 4 | 0 | Electronic billing | {0, No}... | None | 6 | Right | Nominal | Input |
| 25 | custcat | Numeric | 8 | 0 | Customer category | {1, Basic service}... | None | 10 | Right | Nominal | Input |
| 26 | churn | Numeric | 4 | 0 | Switched provider within last month | {0, No}... | None | 6 | Right | Nominal | Input |
| 27 | | | | | | | | | | | |
| 28 | | | | | | | | | | | |
| 29 | | | | | | | | | | | |


Output
Classification Tree
Title
Notes
Active Dataset
Warnings
Model Summary
Tree Diagram
Risk
Classification
Validation None
Maximum Tree Depth 3
Minimum Cases in Parent Node 100
Minimum Cases in Child Node 50
Results
Independent Variables Included Months with service, Internet, Equipment rental
Number of Nodes 15
Number of Terminal Nodes 10
Depth 2
Risk
Estimate Std. Error
.229 .013
Growing Method: CHAID
Dependent Variable: Switched provider within last month


Output
Classification Tree
Title
Notes
Active Dataset
Warnings
Model Summary
Tree Diagram
Risk
Classification
Validation
None
Maximum Tree Depth
3
Minimum Cases in Parent Node
100
Minimum Cases in Child Node
50
Results
Independent Variables Included
Months with service, Internet, Equipment rental
Number of Nodes
15
Number of Terminal Nodes
10
Depth
2
Tree Editor
File View Rules Options Help
Risk
Estimate
Std. Error
.229
.013
Growing Method: CHAID
Dependent Variable: Switched provider within last month


Output
Classification Tree
Title
Notes
Active Dataset
Warnings
Model Summary
Tree Diagram
Risk
Classification
Validation
None
Maximum Tree Depth
3
Minimum Cases in Parent Node
100
Minimum Cases in Child Node
50
Results
Independent Variables Included
Months with service, Internet, Equipment rental
Number of Nodes
15
Number of Terminal Nodes
10
Depth
2
Tree Editor
File
View
Rules
Options
Help
Risk
Estimate
Std. Error
.229
.013
Growing Method: CHAID
Dependent Variable: Switched provider within last month


Output
Classification Tree
Title
Notes
Active Dataset
Warnings
Model Summary
Tree Diagram
Risk
Classification
Internet
Adj. P-value=0.015, Chi-square=5.990, df=1
Equipment rental
Adj. P-value=0.000, Chi-square=24.987, df=1
Equipment rental
Adj. P-value=0.000, Chi-square=20.631, df=1
Equipment rental
Adj. P-value=0.001, Chi-square=11.964, df=1
No
Yes
Node 7
Category % n
No 63.5 33
Yes 36.5 19
Total 5.2 52
Node 8
Category % n
No 40.0 22
Yes 60.0 33
Total 5.5 55
Node 9
Category % n
No 78.9 90
Yes 21.1 24
Total 11.4 114
Node 10
Category % n
No 44.8 39
Yes 55.2 48
Total 8.7 87
Node 11
Category % n
No 85.8 175
Yes 14.2 29
Total 20.4 204
Node 12
Category % n
No 63.1 65
Yes 36.9 38
Total 10.3 103
Node 13
Category % n
No 94.4 119
Yes 5.6 7
Total 12.6 126
Risk
Estimate Std. Error
.229 .013
Growing Method: CHAID
Dependent Variable: Switched provider within last month
Classification
Predicted
Observed No Yes Percent Correct
No 631 95 86.9%
Yes 134 140 51.1%
Overall Percentage 76.5% 23.5% 77.1%
Growing Method: CHAID
Dependent Variable: Switched provider within last month

Σας ευχαριστώ!