

Ατομική Διπλωματική Εργασία

**ΕΥΦΥΕΣ ΣΥΣΤΗΜΑ ΑΞΙΟΛΟΓΗΣΗΣ ΤΗΣ
ΑΞΙΟΠΙΣΤΙΑΣ ΤΩΝ ΠΕΛΑΤΩΝ ΜΕ ΤΗ ΧΡΗΣΗ
ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ ΚΑΙ
ΓΕΝΕΤΙΚΩΝ ΑΛΓΟΡΙΘΜΩΝ**

Χρίστος Λουκά

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ



ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάϊος 2009

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

**Ευφύες σύστημα αξιολόγησης της αξιοπιστίας των πελατών με τη χρήση
νευρωνικών δικτύων και γενετικών αλγορίθμων**

Χρίστος Λουκά

Επιβλέπων Καθηγητής

Ανδρέας Ανδρέου

Η Ατομική Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση των
απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του Τμήματος Πληροφορικής του
Πανεπιστημίου Κύπρου

Μάϊος 2009

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή κύριο Ανδρέα Ανδρέου και τον ερευνητή κύριο Κωνσταντίνο Στυλιανού, για την καθοδήγηση και τη βοήθεια που μου πρόσφεραν κατά τη διάρκεια της υλοποίησης της διπλωματικής μου εργασίας, αλλά και για την επίτευξη μιας άψογης συνεργασίας.

Περίληψη

Το σύστημα που υλοποίησα, χρησιμοποιεί έξυπνες μεθόδους και έχει ως στόχο την επίτευξη της μέγιστης δυνατής επιτυχίας πρόβλεψης, όσον αφορά την αξιοπιστία ενός πελάτη, να ξεπληρώσει την τράπεζα για ένα δάνειο που του δόθηκε. Συγκεκριμένα υλοποίησα 3 διαφορετικές μεθόδους, οι οποίες στοχεύουν τόσο στην επίτευξη καλού αποτελέσματος, όσο και στην εξεύρεση των σημαντικότερων παραμέτρων του δικτύου αλλά και των σημαντικότερων παραγόντων που επηρεάζουν την εκπαίδευση του δικτύου, σύμφωνα πάντα με το ανάλογο dataset που χρησιμοποιείται. Οι 3 αυτοί μέθοδοι χρησιμοποιούν τον αλγόριθμο Back – Propagation, γενετικούς αλγόριθμους και την τεχνική Inputs Sensitivity Analysis αντίστοιχα. Σε γενικές γραμμές κατάφερα να πετύχω αρκετά καλά αποτελέσματα, με ψηλότερο το ποσοστό επιτυχίας της τάξεως του 72%. Σε σύγκριση με το άρθρο 'Credit scoring using the hybrid neural discriminant technique' των Tian-Shyug Lee, Chih-Chou Chiu, Chi-Jie Lu, και I-Fei Chend πιστεύω τα αποτελέσματα μου ήταν πολύ καλά, αφού οι συγγραφείς στην περίπτωση τους κατάφεραν με απλό νευρωνικό δίκτυο να πετύχουν αποτελέσματα μέχρι και 73,7%, όχι πολύ πιο πάνω από τα δικά μου. Κατάφερα να φτάσω στο συμπέρασμα ότι τα κανονικοποιημένα δεδομένα μπορούν να είναι πιο αποτελεσματικά, όσο αφορά την εκπαίδευση του δικτύου. Ακόμη κατάφερα να δείξω ότι ο συνδυασμός συγκεκριμένων παραμέτρων του δικτύου και κάποιων γνωρισμάτων μπορεί να οδηγήσει σε καλύτερα αποτελέσματα. Η δεύτερη μέθοδος συγκεκριμένα μπορεί να δώσει αρκετούς συνδυασμούς για την επίτευξη όσο το δυνατό βέλτιστων αποτελεσμάτων. Τέλος κατάφερα να απομονώσω κάποια από τα γνωρίσματα, τα οποία έχουν συνήθως το μεγαλύτερο ποσοστό συμβολής στην εκπαίδευση του δικτύου και παίζουν σημαντικότερο ρόλο από τα άλλα. Τα γνωρίσματα αυτά είναι:

- Σκοπός δανείου
- Ποσό δανείου
- Διάρκεια
- Πληρωμή προηγούμενων δανείων
- Κατάσταση λογαριασμού
- Περιουσιακά στοιχεία

Περιεχόμενα

Κεφάλαιο 1	Εισαγωγή.....	1
	1.1 Ορισμός του Credit scoring	1
	1.2 Πρόβλημα που καλείται να λύσει το Credit scoring	1
	1.3 Σημαντικότητα	2
Κεφάλαιο 2	Credit Scoring.....	3
	2.1 Ιστορικό	3
	2.2 Θεωρητικό υπόβαθρο	4
Κεφάλαιο 3	Τεχνικό υπόβαθρο.....	8
	3.1 Τεχνητά νευρωνικά δίκτυα	8
	3.2 Αλγόριθμος Back propagation	11
	3.3 Εξελικτικοί αλγόριθμοι	14
	3.4 Γενετικοί αλγόριθμοι	15
	3.5 Inputs Sensitivity Analysis	17
Κεφάλαιο 4	Μεθοδολογία.....	18
	4.1 Σκοπός συστήματος	18
	4.2 Dataset που χρησιμοποιήθηκε	18
	4.3 Επεξεργασία δεδομένων	19
	4.4 Πρώτη μέθοδος: Νευρωνικό δίκτυο με αλγόριθμο Back – Propagation	20
	4.5 Δεύτερη μέθοδος: Γενετικοί αλγόριθμοι	32
	4.6 Τρίτη μέθοδος: Inputs Sensitivity Analysis	37
	4.7 Ενσωμάτωση animation στο σύστημα	42
Κεφάλαιο 5	Αποτελέσματα	44
	5.1 Αποτελέσματα 1 ^{ης} μεθόδου	44
	5.2 Αποτελέσματα 2 ^{ης} μεθόδου	54

5.3 Αποτελέσματα 3 ^{ης} μεθόδου	57
Κεφάλαιο 6 Συμπεράσματα	60
6.1 Γενικά συμπεράσματα	60
6.2 Μελλοντική εργασία	60
Βιβλιογραφία	62
Παράρτημα Α	A-1

Κεφάλαιο 1

Εισαγωγή

1.1 Ορισμός του Credit Scoring	1
1.2 Πρόβλημα που καλείται να λύσει το Credit Scoring	1
1.3 Σημαντικότητα	2

1.1 Ορισμός του Credit Scoring

Το credit scoring είναι μια τεχνική η οποία βασίζεται σε στατιστικές και έξυπνες μεθόδους και χρησιμοποιείται για την εξακρίβωση της αξιοπιστίας των δανειζόμενων. Ανάλογα με το ποσοστό αξιοπιστίας τους, οι δανειστές μπορούν να αποφασίσουν το χρηματικό ποσό που θα τους δανείσουν. Η τεχνική αυτή θεωρείται αρκετά αξιόπιστη, αφού χρησιμοποιεί πραγματικά δεδομένα από διάφορους πελάτες και προσπαθεί να τα αναλύσει με στόχο την αναγνώριση των παραγόντων που συμβάλλουν περισσότερο στην αξιοπιστία των δανειζόμενων.

1.2 Πρόβλημα που καλείται να λύσει το Credit Scoring

Εδώ και αρκετά χρόνια διάφοροι οργανισμοί και εταιρείες, στην πλειοψηφία τους τράπεζες, προσπαθούν να υιοθετήσουν διάφορες αυτοματοποιημένες τεχνικές, οι οποίες θα τους βοηθήσουν τόσο στις αποφάσεις τους σε θέματα έγκρισης δανείων, όσο και στην εξοικονόμηση χρόνου και κόστους. Το μεγαλύτερο πρόβλημα που υπάρχει εδώ και χρόνια, είναι τα χρέη που δημιουργούνται στους χρηματοοικονομικούς οργανισμούς, λόγω της αδυναμίας διάφορων πελατών να ξεπληρώσουν τα δάνεια που τους παραχωρήθηκαν. Παράδειγμα αποτελεί το πρόσφατο θέμα χρεωκοπίας δύο μεγάλων τραπεζών της Αμερικής, που οφειλόταν στη συνεχή και απερίσκεπτη

παραχώρηση δανείων, χωρίς την απαιτούμενη μελέτη, όσον αφορά την αξιοπιστία και δυνατότητα των πελατών να ξεπληρώσουν το δανειζόμενο ποσό. Για το σκοπό αυτό έχουν αναπτυχθεί διάφορα συστήματα τα οποία διαχωρίζουν τους πελάτες ως κακοπληρωτές ή καλοπληρωτές, βασισμένα σε παλαιότερες συναλλαγές τους. Με τον τρόπο αυτό, τα συστήματα παρέχουν περισσότερες πληροφορίες στους αρμόδιους για τα θέματα δανεισμού, βελτιώνοντας έτσι την ορθότητα των αποφάσεων τους.

1.3 Σημαντικότητα

Το Credit Scoring αυτό είναι πολύ σημαντικό, αφού συντελεί στη μείωση του ρίσκου που παίρνει ένας χρηματοοικονομικός οργανισμός με το να δανείσει χρήματα, μιας και οι πιθανότητες να του επιστραφούν πίσω είναι μεγαλύτερες, λόγω των ορθότερων αποφάσεων. Ειδικά την περίοδο αυτή, που επικρατεί παγκόσμια οικονομική κρίση, κάθε τράπεζα έχει την ανάγκη να κάνει προσεκτικά βήματα και επιλογές. Το Credit Scoring είναι μια μέθοδος η οποία βοηθά πάρα πολύ, ειδικά στον τομέα της ορθότητας των αποφάσεων αλλά και στην μείωση ζημιών. Ένας από τους κύριους λόγους χρησιμοποίησης του είναι η βελτίωση της λειτουργικής αποτελεσματικότητας καθώς και ο καλύτερος έλεγχος του πορτοφολίου. Έχει γίνει ήδη ένα μέσο εσωτερικής επικοινωνίας, δημιουργώντας ένα μηχανισμό ο οποίος μπορεί να μετατρέψει την πολιτική ενός οργανισμού σε πράξη, παρέχοντας εύκολα πληροφορίες για κάθε δανειζόμενο και απλουστεύοντας τη λήψη αποφάσεων.

Κεφάλαιο 2

Credit Scoring

2.1 Ιστορικό	3
2.2 Θεωρητικό υπόβαθρο	4

2.1 Ιστορικό

Τα σημερινά συστήματα για Credit scoring δουλεύουν σχεδόν ασταμάτητα, για να παρέχουν σε διάφορους δανειστές και εργαζόμενους τις πληροφορίες που χρειάζονται, για να προχωρήσουν στη λήψη αποφάσεων. Ωστόσο τα συστήματα αυτά δεν λειτουργούσαν πάντοτε τόσο απλά και αποτελεσματικά όπως σήμερα. Οι διάφορες τεχνικές Credit scoring του σήμερα, έχουν εξελιχθεί από απλές τοπικές συναλλαγές πληροφοριών οι οποίες γίνονταν μόνο μεταξύ των ενδιαφερόμενων, σε ένα σύστημα το οποίο ωφελεί όλες τις εμπλεκόμενες πλευρές. Το Credit scoring γεννήθηκε κατά τη διάρκεια του 19^{ου} αιώνα, όταν μικροί έμποροι λιανικής, εμπορεύονταν οικονομικές πληροφορίες οι οποίες αφορούσαν τους πελάτες τους. Αυτοί οι συνεταιρισμοί, οργανώθηκαν στο τέλος σε μικρά credit bureaus και αποτελούσαν ένα ανεπίσημο σύστημα ανταλλαγής πληροφοριών μεταξύ των εμπόρων. Στις αρχές του 20^{ου} αιώνα έκαναν την εμφάνιση τους εφευρέσεις της μοντέρνας εποχής όπως το αυτοκίνητο και το τηλέφωνο. Οι εφευρέσεις αυτές καθώς εξαπλώνονταν, ο κόσμος γινόταν ολοένα και πιο ‘μικρός’. Οι άνθρωποι μπορούσαν να ταξιδέψουν όπουδήποτε για να κλείσουν συμφωνίες και δουλειές, ενώ οι δανειστές μπορούσαν εύκολα να μοιραστούν πληροφορίες για ευρύτερες περιοχές. Με την συνεχή ανάπτυξη της τεχνολογίας τα μικρά αυτά credit bureaus, άρχισαν να στερεοποιούνται και να μετατρέπονται σε μεγάλους και έμπειρους τώρα ομίλους. Η έλευση των ηλεκτρονικών υπολογιστών σηματοδότησε τη μοντέρνα εποχή για τους ομίλους αυτούς, επιτρέποντας την αποθήκευση τεράστιων όγκων δεδομένων και την εύκολη και γρήγορη ανταλλαγή πληροφοριών, με τρόπους που ήταν αδύνατοι παλιά. Οι όμιλοι αυτοί άρχισαν να

συγχωνεύονται και να μεγαλώνουν σχηματίζοντας έτσι τους ‘γίγαντες’ που υπάρχουν σήμερα με τις Equifax, Trans Union and Experian να είναι από τις πιο γνωστές στις Η.Π.Α. Όσο τα credit bureaus μεγάλωναν, τόσο μεγάλωνε και η σημασία τους στη ζωή του μέσου πολίτη. Το γεγονός αυτό είχε ως αποτέλεσμα την αύξηση της σημαντικότητας των δεδομένων που συλλέγονταν αλλά και τον τρόπο που χρησιμοποιούνταν. Τα σημερινά συστήματα Credit scoring έχουν εξελιχθεί με πρότυπα την κοινωνία μας και το μοντέρνο οικονομικό σύστημα. Καθώς η ανάπτυξη των μοντέρνων συγκοινωνιών και επικοινωνιών επηρέαζε τον τόπο και τρόπο με τον οποίο οι άνθρωποι διεξήγαγαν τις δουλειές τους, τα credit bureaus ανταποκρίθηκαν με το να αυξήσουν την εμβέλεια τους. Τέλος, με την εποχή της πληροφορίας να γίνεται πραγματικότητα, το credit scoring κατάφερε να εξελιχθεί σε ένα σύστημα το οποίο είναι ωφέλιμο όχι μόνο για τους δανειστές, αλλά και για τους ίδιους τους καταναλωτές.

2.2 Θεωρητικό υπόβαθρο

Για την υλοποίηση συστημάτων Credit Scoring, χρησιμοποιούνται πολλές τεχνικές. Μια από τις συχνότερες, είναι αυτή των νευρωνικών δικτύων η οποία υπόσχεται αρκετά καλά αποτελέσματα και συγκεκριμένα της τάξεως του 70%. Ένα από τα μειονεκτήματα της τεχνικής αυτής όμως είναι το γεγονός ότι είναι αρκετά χρονοβόρα. Άλλες κοινές τεχνικές οι οποίες χρησιμοποιούνται ήδη, είναι οι στατιστικές μέθοδοι Logistic regression και Discriminant analysis, οι οποίες είναι αρκετά αξιόπιστες. Χρησιμοποιήθηκαν επίσης και άλλες τεχνικές όπως τα decision trees, γενετικοί αλγόριθμοι, γενετικός προγραμματισμός καθώς και data mining με λιγότερο όμως ποσοστό επιτυχίας στα αποτελέσματα τους. Χρησιμοποιήθηκαν επίσης και υβριδικά συστήματα τα οποία όμως βρίσκονται ακόμη σε πειραματικό στάδιο. Κατά τη διάρκεια του 1^{ου} εξαμήνου μελέτησα κάποια άρθρα τα οποία αναφέρονται στο θέμα του Credit scoring, με στόχο την καλύτερη κατανόηση του θέματος αλλά και τη μελέτη διάφορων ιδεολογιών, όσο αφορά τον τρόπο που θα μπορούσα να υλοποιήσω το σύστημα μου.

Οι Tian-Shyug Lee et al το 2002 [1], χρησιμοποίησαν ένα σύνολο δεδομένων μιας πιστωτικής κάρτας, από μια τοπική τράπεζα στην Ταϊβάν. Για κάθε πελάτη, υπάρχουν εννέα μεταβλητές οι οποίες είναι: φύλο, ηλικία, κατάσταση γάμου, επίπεδο εκπαίδευσης, επάγγελμα, θέση εργασίας, ετήσιο εισόδημα, οικιστική κατάσταση και

όρια πίστωσης. Η μεταβλητή εξόδου θα είναι η κατάσταση πίστωσης του, καλή ή κακή. Επέλεξαν τυχαία 6000 σύνολα δεδομένων, από τα οποία 4000 θα χρησιμοποιήσουν ως δεδομένα εκπαίδευσης και τα υπόλοιπα 2000 ως δεδομένα επαλήθευσης. Το δίκτυο που χρησιμοποίησαν είναι τύπου feedforward και συγκεκριμένα αυτό που χρησιμοποιεί τον αλγόριθμο Back Propagation. Χρησιμοποίησαν ένα κρυφό επίπεδο μόνο και πειραματίστηκαν όσον αφορά τον αριθμό των κρυφών νευρώνων και το ρυθμό μάθησης. Οι εννέα μεταβλητές που ανέφερα πιο πάνω ήταν οι εισοδοί του δικτύου. Άφησαν το δίκτυο να εκπαιδευθεί για 3000 επαναλήψεις. Κατέληξαν ότι την καλύτερη απόδοση και ακρίβεια είχε το δίκτυο με 19 κρυμμένους νευρώνες και με ρυθμό μάθησης 0.004, αφού είχε το μικρότερο σφάλμα (RMSE). Τα αποτελέσματα τους ήταν αρκετά ικανοποιητικά με μέσο ορθό ποσοστό κατηγοριοποίησης 73.70% και συγκεκριμένα 85.27% για τους καλοπληρωτές και 62.34% για τους κακοπληρωτές. Παρατήρησαν ότι όσον αφορά τους κακοπληρωτές, το ποσοστό λανθασμένης κατηγοριοποίησης είναι πολύ μεγαλύτερο από αυτό των καλοπληρωτών. Στη συνέχεια οι συγγραφείς δοκίμασαν ένα υβριδικό μοντέλο, χρησιμοποιώντας το ίδιο νευρωνικό δίκτυο με πιο πάνω, με τη διαφορά ότι οι εισοδοί θα ήταν επτά και συγκεκριμένα οι έξι μεταβλητές που δέχεται το μοντέλο LDA (στατιστική μέθοδος), μαζί με την έξοδο του. Οι έξι μεταβλητές είναι ετήσιο εισόδημα, οικιστική κατάσταση, όρια πίστωσης, φύλο, ηλικία και επάγγελμα. Στόχος τους ήταν να μελετήσουν το γεγονός εάν τα αποτελέσματα θα ήταν βελτιωμένα, αφού ουσιαστικά θα επεξεργάζονταν από δύο συστήματα. Τα αποτελέσματα ήταν βελτιωμένα σε σχέση με τα προηγούμενα. Συγκεκριμένα το μέσο ορθό ποσοστό κατηγοριοποίησης ήταν 77%. Το πρόβλημα όμως του μεγάλου λανθασμένου ποσοστού κακοπληρωτών, παραμένει και είναι σε αυτήν την περίπτωση 27.45%.

Οι Li zhan et al το 2004 [3], χρησιμοποίησαν 2 μεθόδους για την εξαγωγή των αποτελεσμάτων τους και συγκεκριμένα γενετικούς αλγόριθμους και νευρωνικά δίκτυα. Χρησιμοποίησαν 18 χαρακτηριστικά πελατών ως εισόδους, χωρίς όμως να τα αναφέρουν. Το δίκτυο που χρησιμοποίησαν εφάρμοζε τον αλγόριθμο Back Propagation και είχε ως έξοδο το score του πελάτη. Εάν αυτό ήταν πάνω από μια τιμή, τότε χαρακτηρίζονταν ως 'good credit', αλλιώς ως 'bad credit'. Χρησιμοποίησαν ένα κρυφό επίπεδο με 10 κρυφούς νευρώνες. Τα δεδομένα που χρησιμοποίησαν ήταν από μια εταιρεία τηλεπικοινωνιών στην Κίνα, με την ονομασία 'Go tone'. Για την εκπαίδευση

του δικτύου χρησιμοποιήθηκαν 20000 πελάτες, που αντιστοιχούν στο 5% των πελατών της εταιρείας σε ένα υποκατάστημα. Τα αποτελέσματα ήταν αρκετά καλά. Συγκεκριμένα μόνο 11% από τους κακοπληρωτές κατηγοριοποιήθηκε στην αντίθετη κατηγορία, ενώ για τους καλοπληρωτές το 13%. Να αναφέρω ότι οι συγγραφείς προβάλλουν κάποια πλεονεκτήματα των γενετικών αλγόριθμων, λέγοντας ότι είναι πιο αποδοτικοί και βέλτιστοι στην επεξεργασία, ειδικά σε περιπτώσεις μεγάλων όγκων δεδομένων.

Οι M. B. Yobas et al (2000) [2], χρησιμοποίησαν δεδομένα από 1001 ανθρώπους οι οποίοι είχαν πιστωτική κάρτα, την είχαν χρησιμοποιήσει και το ιστορικό των συναλλαγών τους ήταν γνωστό. Σε περίπτωση που κάποιος από αυτούς είχε χάσει μια ή περισσότερες πληρωμές για εξόφληση του χρέους του, τότε αυτή χαρακτηριζόταν ως 'slow', αλλιώς ως 'good'. Το κριτήριο αυτό ίσως είναι λίγο αυστηρό φυσικά, αφού για να 'στιγματιστεί' κάποιος πελάτης από την τράπεζα ως κακοπληρωτής, θα έπρεπε να μην πληρώσει 3 ή περισσότερες δόσεις. Χρησιμοποίησαν 14 χαρακτηριστικά, τα οποία είναι: εργασιακή κατάσταση του αιτητή, χρόνια στην τράπεζα ως πελάτης, αξία υποθήκης σπιτιού, αριθμός παιδιών, χρόνια εργοδότησης στο υφιστάμενο επάγγελμα, κατοικία, τύποι λογαριασμών, κάρτες στην κατοχή του, αξία σπιτιού, τηλέφωνο οικίας, εισόδημα, εισόδημα συζύγου. Ορισμένες αλφαβητικές μεταβλητές μετατράπηκαν σε αριθμητικές για να εισαχθούν στο σύστημα. Για παράδειγμα εάν μια μεταβλητή είχε 6 αλφαβητικές αξίες, τότε γραφόταν σαν ένας αριθμός μεταξύ του 0 και 5. Χρησιμοποίησαν διάφορους ρυθμούς μάθησης, ένα ή 2 κρυφά επίπεδα, 23-29 εισόδους, 18-5 κρυφούς νευρώνες για το πρώτο κρυφό επίπεδο και 5-3 για το δεύτερο. Υπήρχαν επίσης 2 νευρώνες εξόδου. Ο αριθμός των επαναλήψεων κυμάνθηκε από 40000 ως 200000. Το δίκτυο εκπαιδεύθηκε στις 501 περιπτώσεις και επαληθεύτηκε χρησιμοποιώντας τις υπόλοιπες 500. Τα αποτελέσματα ήταν και σε αυτήν την περίπτωση ικανοποιητικά όσον αφορά τους καλοπληρωτές με 79% ορθό ποσοστό κατηγοριοποίησης. Δεν ισχύει το ίδιο όμως για τους κακοπληρωτές αφού μόνο το 39% αναγνωρίστηκε ορθά.

Σύμφωνα με τις πιο πάνω μελέτες, μπορεί κάποιος να παρατηρήσει εύκολα, ότι τα αποτελέσματα δεν είναι αρκετά ικανοποιητικά σε βαθμό που να καθιστούν τις μεθόδους αυτούς εντελώς αξιόπιστες. Επομένως υπάρχει ακόμη αρκετός χώρος για

ανάπτυξη του συγκεκριμένου θέματος. Κύριος στόχος μου θα είναι η ανάπτυξη βελτιωμένων μεθόδων και η προσπάθεια επίτευξης αυξημένων ποσοστών επιτυχίας.

Κεφάλαιο 3

Τεχνικό υπόβαθρο

3.1 Τεχνητά νευρωνικά δίκτυα	8
3.2 Αλγόριθμος Back Propagation	11
3.3 Εξελικτικοί αλγόριθμοι	14
3.4 Γενετικοί αλγόριθμοι	15
3.5 Inputs Sensitivity Analysis	17

3.1 Τεχνητά νευρωνικά δίκτυα

Η ανάγκη του ανθρώπου να μοντελοποιήσει το σημαντικότερο του όργανο που δεν είναι άλλο φυσικά από τον ανθρώπινο εγκέφαλο, και το πώς αυτός λειτουργεί οδήγησε στη ανάγκη για τη δημιουργία των τεχνητών νευρωνικών δικτύων. Με τον όρο τεχνητό νευρωνικό δίκτυο αναφερόμαστε σε μια διασυνδεμένη ομάδα τεχνητών νευρώνων που χρησιμοποιώντας ένα μαθηματικό ή υπολογιστικό μοντέλο επεξεργάζεται πληροφορίες βασισμένη σε μια πολυσύνδετη προσέγγιση ενός προβλήματος και παράγει κάποια συγκεκριμένη έξοδο. Όπως και τα νευρωνικά δίκτυα του ανθρώπινου εγκεφάλου, τα τεχνητά νευρωνικά δίκτυα έχουν την ικανότητα να μαθαίνουν μέσω παραδειγμάτων, να εκπαιδεύονται και κατά συνέπεια να αντιδρούν ορθότερα σε οποιαδήποτε μελλοντική είσοδο δεχτούν. Στις περισσότερες των περιπτώσεων ένα νευρωνικό δίκτυο παρουσιάζεται ουσιαστικά σαν ένα προσαρμοστικό σύστημα που αλλάζει τη δομή του βασισμένο στις εξωτερικές ή εσωτερικές πληροφορίες που το διατρέχουν. Σε πρακτικότερους όρους τα νευρωνικά δίκτυα μπορούν να χαρακτηριστούν σαν μη γραμμικά εργαλεία διαμόρφωσης στατιστικών στοιχείων. Μπορούν να χρησιμοποιηθούν για να μοντελοποιήσουν σύνθετες σχέσεις μεταξύ εισόδων και εξόδων ή για τη ομαδοποίηση προτύπων δεδομένων. Σε αρκετές περιπτώσεις και προ πολύπλοκων προβλημάτων όπου η μοντελοποίηση ή επίλυση τους με ορθόδοξα μαθηματικά μοντέλα φάνταζε αδύνατη, η μοντελοποίηση τους με τεχνητά νευρωνικά

δίκτυα ήταν η καλύτερη και πιο προσεγγιστική στο αποτέλεσμα λύση. Η χρήση των τεχνητών νευρωνικών δικτύων σήμερα βρίσκει πολλές εφαρμογές σε αρκετούς κλάδους των φυσικών επιστημών όπως είναι η πληροφορική, τα μαθηματικά, η στατιστική και η φυσική, σε αρκετούς κλάδους της ιατρικής όπως είναι η ψυχολογία, η βιολογία και η νευρολογία καθώς επίσης και κλάδους των οικονομικών επιστημών όπως είναι οι επιχειρήσεις και τα οικονομικά.

Κατηγορίες νευρωνικών δικτύων:

Υπάρχουν τρεις σημαντικές κατηγορίες μάθησης με τις οποίες μπορεί να εκπαιδευτεί ένα νευρωνικό δίκτυο έτσι ώστε να φτάσει στην επίλυση ενός συγκεκριμένου προβλήματος. Η φύση του προβλήματος καθώς επίσης και η γνώση των δεδομένων που έχουμε για το πρόβλημα αυτό μας οδηγούν στο ποια είναι η καταλληλότερη να χρησιμοποιηθεί. Πολλές φορές βέβαια ένα πρόβλημα μπορεί να προσεγγιστεί από διαφορετικές κατηγορίες μάθησης. Οι τρεις αυτές κυριότερες κατηγορίες μάθησης που χρησιμοποιούνται για την εκπαίδευση τεχνητών νευρωνικών δικτύων είναι:

i) Επιβλεπόμενη Μάθηση:

Μάθηση που για να χρησιμοποιηθεί για εκπαίδευση ενός δικτύου χρειάζεται να υπάρχει γνώση τόσο των διανυσμάτων εισόδου αλλά και των επιθυμητών εξόδου για το καθένα από αυτό. Αποκαλείται συχνά μάθηση με δάσκαλο που λέει ουσιαστικά στο δίκτυο μας αυτό που είπαμε πιο πάνω, ποια είναι δηλαδή η απαιτούμενη έξοδος για κάθε δεδομένη είσοδο. Τα βάρη του δικτύου μεταβάλλονται με σκοπό να μειώσουν όσο το δυνατότερο γίνεται τη διαφορά μεταξύ της επιθυμητής και της πραγματικής εξόδου για κάθε πρότυπο εισόδου και κατ' επέκταση τη συνάρτηση λάθους του δικτύου. Οι κυριότεροι αλγόριθμοι επιβλεπόμενης μάθησης είναι οι Delta και Back Propagation Rule.

ii) Μη Επιβλεπόμενη Μάθηση:

Μη επιβλεπόμενη μάθηση αφού δεν υπάρχει γνώση των επιθυμητών εξόδων που παράγονται από το δίκτυο. Με τη χρήση της μάθησης αυτής ζητείται ουσιαστικά από το δίκτυο να κοιτάξει για εύρεση παρόμοιων χαρακτηριστικών ή στατιστικών ομοιοτήτων

στο σύνολο των εισόδων που δέχεται και να κάνει με τον δικό του τρόπο ένα είδος κατηγοριοποίησης των πιο πάνω εισόδων εκπαίδευσης. Κάθε μέλος τώρα, σε οποιαδήποτε ομάδα και να ανήκει από αυτέςπου θα σχηματιστούν από το δίκτυο, είναι ουσιαστικά ένα διάνυσμα εισόδου που έχει παρόμοια χαρακτηριστικά και με τα υπόλοιπα μέλη της κατηγορίας αυτής. Στη μη επιβλεπόμενη μάθηση το δίκτυο δημιουργεί τις γειτονιές του αρχικά βασισμένο στα πιο πάνω και σε μια αρχική ακτίνα γειτονιάς. Στην πορεία η ακτίνα αυτή μειώνεται έτσι ώστε να συμμαζευτεί ουσιαστικά η γειτονιά. Οι κυριότεροι αλγόριθμοι μη επιβλεπόμενης μάθησης είναι οι Hebbian Learning και Kohonen's self-organizing map.

iii) Ενισχυτική Μάθηση:

Στο συγκεκριμένο είδος μάθησης το δίκτυο μετά από κάθε έξοδο του, λαμβάνει ένα σήμα από το περιβάλλον το οποίο μπορεί να είναι είτε θετικό (αμοιβή) είτε αρνητικό (πέναλτι). Τα βάρη του δικτύου αλλάζουν έτσι ώστε να αναπτύξουν μια συμπεριφορά εισόδου/εξόδου η οποία μεγιστοποιεί την πιθανότητα, το σήμα που λαμβάνεται από το περιβάλλον να είναι αμοιβή και ελαχιστοποιεί την πιθανότητα αυτό να είναι πέναλτι. Με τον ίδιο τρόπο που θεωρήσαμε πριν, ότι η επιβλεπόμενη μάθηση είναι μάθηση με δάσκαλο, έτσι και η ενισχυτική μάθηση μπορούμε να πούμε ότι είναι μάθηση με κριτή ο οποίος στέλνει στο σύστημα τα σήματα που αναφέραμε πιο πάνω. Τονίζεται ότι τα σήματα που επιστρέφονται από το περιβάλλον στο σύστημα είναι μόνο αξιολογικά και σε καμία περίπτωση καθοδηγητικά αφού ο ίδιος ο μαθητευόμενος πρέπει να ανακαλύψει μέσα από εξερεύνηση ποιες ενέργειες επιστρέφουν την μεγαλύτερη αμοιβή (γνώση που αποκτείται με την πείρα). Ένα κλασσικό παράδειγμα ενισχυτικής μάθησης είναι ο Barto's associative reward-penalty rule.

Πλεονεκτήματα νευρωνικών δικτύων:

Σημαντικά προτερήματα των νευρωνικών δικτύων έναντι των τυποποιημένων υπολογιστικών μοντέλων είναι:

- i) Ταχύτητα λόγω του παραλληλισμού σε αντίθεση της σειριακής φύσης των τυποποιημένων υπολογιστών.
- ii) Ανθεκτικότητα στα σφάλματα αφού η απώλεια ενός κόμβου-νευρώνα σε ένα δίκτυο

δεν σταματά την εκτέλεση της λειτουργίας του και της αποδοτικότητας του σε αντίθεση με την ευθραυστότητα των τυποποιημένων υπολογιστών.

- iii) Προσαρμογή και μάθηση μέσω παραδειγμάτων.
- iv) Δυνατότητα γενίκευσης μέσα από προηγούμενα παραδείγματα.
- v) Μετά την εκπαίδευση οι υπολογιστικές του απαιτήσεις είναι πάρα πολύ μικρές.

3.2 Αλγόριθμος Back Propagation

Πρόκειται για το σημαντικότερο και πιο διαδεδομένο αλγόριθμο μάθησης με επίβλεψη που χρησιμοποιείται για εκπαίδευση νευρωνικών δικτύων. Όπως αναφέρθηκε και πιο πάνω στην επιβλεπόμενη μάθηση το δίκτυο μας μαθαίνει από παραδείγματα. Βάση αυτών των παραδειγμάτων λοιπόν, αναλύοντας σε κάθε επανάληψη τα λάθη στους νευρώνες εξόδου αναπροσαρμόζει τα βάρη του μέχρι ούτως φτάσει στο επιθυμητό αποτέλεσμα. Ο αλγόριθμος αυτός ανακαλύφθηκε το 1974 από τον Paul Werbos αλλά δεν είχε την ευρεία και δημοσιότητα παρά μόνο περίμενε να γίνει αυτό το 1986 από τους David E. Rumelhart, Geoffrey E. Hinton και Ronald J. Williams. Ο αλγόριθμος αυτός χρησιμοποιείται και είναι αποδοτικός στην εκπαίδευση feed-forward δικτύων. Με τον όρο feed-forward δίκτυα αναφερόμαστε στα δίκτυα που τα αποτελέσματα από όλους τους νευρώνες πηγαίνουν στα επόμενα αλλά όχι στα προηγούμενα επίπεδα κατά τη διάρκεια της εκπαίδευσης. Ο συγκεκριμένος αλγόριθμος χρησιμοποιείται από το σύστημα μου, όπως θα εξηγήσω στο επόμενο κεφάλαιο.

Μαθηματικές εξισώσεις αλγορίθμου Back-propagation:

$$net_j = \sum_k W_{jk} \cdot O_k \quad (1)$$

Όπου W_{jk} είναι το βάρος που ενώνει το νευρώνα j με το νευρώνα k του προηγούμενου επιπέδου και O_k είναι η τιμή εξόδου του νευρώνα k .

$$O_j = \frac{1}{1 + e^{(-a \cdot net_j)}} \quad (2)$$

Όπου O_j είναι η τιμή εξόδου οποιουδήποτε νευρώνα j όταν το άθροισμα αυτού (net_j) περάσει μέσα από τη σιγμοειδή συνάρτηση.

$$\Delta w_{jk}(t) = \eta \cdot \delta_j O_j + \lambda \cdot \Delta w_{jk}(t-1) \quad (3)$$

Όπου $\Delta w_{jk}(t)$ είναι η αλλαγή στο βάρος w_{jk} στο χρόνο t , η είναι ο ρυθμός μάθησης του δικτύου και λ ο όρος ορμής. Ο υπολογισμός του λάθους δ_j υπολογίζεται διαφορετικά, ανάλογα με τη περίπτωση του κάθε νευρώνα, αν δηλαδή ανήκει στο κρυφό επίπεδο ή στο επίπεδο εξόδου.

$$\delta_i = O_i(1 - O_i)(t_i - O_i) \quad (4)$$

Η εξίσωση (4) υπολογίζει το λάθος στους νευρώνες εξόδου όπου O_i είναι το πραγματικό αποτέλεσμα εξόδου του νευρώνα i και t_i το επιθυμητό αποτέλεσμα του.

$$\delta_i = O_i(1 - O_i) \sum_k W_{jk} \cdot \delta_k \quad (5)$$

Στους νευρώνες του κρυφού επιπέδου που δεν υπάρχει γνώση του επιθυμητού αποτελέσματος ο όρος $t_i - O_i$ αντικαθίσταται από τον όρο $W_{jk} \cdot \delta_k$ όπου δ_k είναι το λάθος στο νευρώνα k του επόμενου επιπέδου και W_{jk} το βάρος που ενώνει το νευρώνα k με το νευρώνα i .

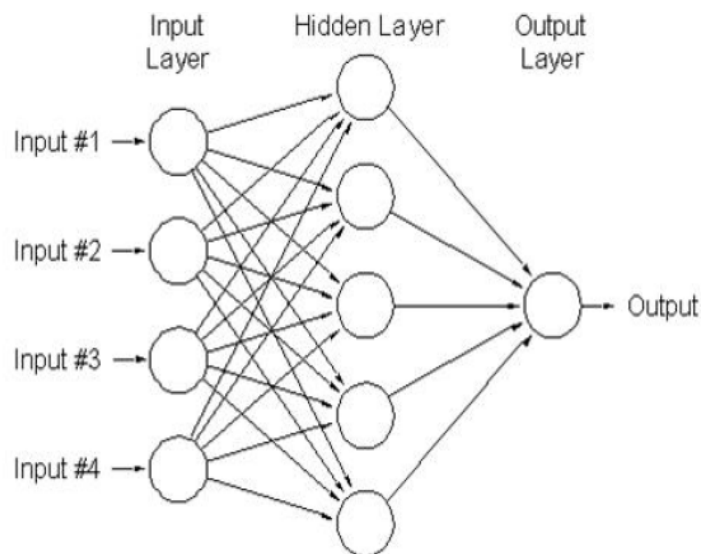
Βήματα του αλγορίθμου μάθησης Back-propagation:

1. Παρουσίαση ενός δεδομένου εκπαίδευσης στο νευρωνικό δίκτυο και αρχικοποίηση των επιθυμητών αποτελεσμάτων στους νευρώνες εξόδου του δικτύου βάσει αυτό.
2. Το δίκτυο υπολογίζει βάσει του συγκεκριμένου παραδείγματος τις πραγματικές τιμές στους νευρώνες εξόδου σύμφωνα με τις εξισώσεις του αλγορίθμου που είδαμε πιο πάνω.
3. Γίνεται σύγκριση μεταξύ των πραγματικών και των επιθυμητών τιμών εξόδου στους νευρώνες εξόδου του δικτύου και υπολογίζεται το σφάλμα του δικτύου.

4. Όπως υποδηλώνει και το όνομα του αλγορίθμου τα λάθη διαδίδονται προς τα πίσω από τους κόμβους παραγωγής στους εσωτερικούς κόμβους και με αυτό τον τρόπο τροποποιούνται τα βάρη των νευρώνων έτσι ώστε να μειωθεί το συνολικό σφάλμα του δικτύου.
5. Η πιο πάνω διαδικασία επαναλαμβάνεται μέχρι ούτως το συνολικό σφάλμα του δικτύου να μειωθεί σε αισθητά χαμηλότερες τιμές.

Παράδειγμα δικτύου:

Στο Σχήμα 3.1 μπορούμε να δούμε ένα μοντέλο νευρωνικού δικτύου BPN. Το δίκτυο είναι χωρισμένο σε επίπεδα. Το πρώτο επίπεδο είναι το επίπεδο εισόδου στο οποίο απλώς οι νευρώνες αρχικοποιούνται με τις τιμές των διανυσμάτων που θα χρησιμοποιηθούν για την εκπαίδευση των δικτύων. Ακολουθώντας έχουμε το κρυφό επίπεδο το οποίο περιλαμβάνει ένα ή δύο στρώματα αφού όπως αποδείχτηκε από το Kolmogorov δύο στρώματα είναι ικανά να λύσουν οποιοδήποτε πρόβλημα όσο αυθαίρετο και αν είναι. Και ακολουθώντας το επίπεδο εξόδου του δικτύου. Ενεργοί νευρώνες του δικτύου είναι όσοι περιλαμβάνονται στο κρυφό ή κρυφά επίπεδα και στο επίπεδο εξόδου του δικτύου. Αποκαλούνται ενεργοί αφού είναι οι νευρώνες που χρησιμοποιούν τη συνάρτηση ενεργοποίησης για τον υπολογισμό της εξόδου τους την εκάστοτε στιγμή. Όλοι οι κόμβοι του δικτύου είναι πλήρως συνδεδεμένοι με τους υπόλοιπους.



Σχήμα 3.1: Παράδειγμα νευρωνικού δικτύου

Προβλήματα αλγορίθμου:

Δύο είναι τα κυριότερα πιθανά προβλήματα που μπορεί να συναντήσει κάποιος κατά την εκπαίδευση ενός δικτύου με τον αλγόριθμο back propagation. Ένα από αυτά είναι το πρόβλημα της υπερβολικής εκπαίδευσης. Το δίκτυο μετά από κάποιο συγκεκριμένο σημείο κατά τη διάρκεια της εκπαίδευσης αντί να γίνεται καλύτερο γίνεται χειρότερο. Ο λόγος που μπορεί να γίνει αυτό είναι γιατί μετά από πολύ εκπαίδευση το δίκτυο ουσιαστικά θα απομνημονεύσει πλέον όλα τα πρότυπα εκπαίδευσης κάτι που είναι απολύτως λάθος αφού ο σκοπός μας είναι το δίκτυο να μάθει μια γενική δομή όσο αφορά τα δεδομένα εισόδου. Οι κυριότεροι τρόποι αντιμετώπισης του προβλήματος αυτού είναι είτε σταματώντας την εκπαίδευση νωρίτερα είτε μειώνοντας την πολυπλοκότητα του δικτύου μειώνοντας τους εσωτερικούς του κόμβους. Το δεύτερο και επίσης πολύ σημαντικό πρόβλημα είναι το πρόβλημα του τοπικού ελαχίστου. Σε μία πιθανή γραφική παράσταση του λάθους σε ένα δίκτυο όπου χρησιμοποιώντας τη μέθοδο κατάβασης κλίσης προσπαθούμε να προσεγγίσουμε το χαμηλότερο της σημείο, υπάρχει η πιθανότητα της παγίδευσης του λάθους σε κάποιο τοπικό ελάχιστο και όχι στο ολικό ελάχιστο της συνάρτησης κάτι που σίγουρα δεν είναι επιθυμητή για εμάς εξέλιξη. Οι κυριότεροι τρόποι αντιμετώπισης του προβλήματος αυτού είναι η μείωση του ρυθμού μάθησης στο δίκτυο (έστω και αν αυτό συνεπάγεται περισσότερο χρόνο για να εκπαιδευτεί), η εισαγωγή του όρου ορμής ο οποίος λαμβάνει υπόψη την αμέσως προηγούμενη αλλαγή των βαρών που έγινε στο δίκτυο (άρα έχουμε μεγάλες αλλαγές στα βάρη αν είναι και οι ήδη υφιστάμενες μεγάλες και το αντίστροφο) και η πρόσθεση εσωτερικών κόμβων στο δίκτυο κάτι που αυξάνει τη δυνατότητα καλύτερης κωδικοποίησης των δεδομένων από το δίκτυο.

3.3 Εξελικτικοί αλγόριθμοι

Στην τεχνητή νοημοσύνη, ένας εξελικτικός αλγόριθμος, είναι ένα υποσύνολο του εξελικτικού υπολογισμού και συγκεκριμένα, ένας αλγόριθμος βελτιστοποίησης βασισμένος στον πληθυσμό του. Ο αλγόριθμος αυτός χρησιμοποιεί ορισμένους μηχανισμούς εμπνευσμένους από τη βιολογική εξέλιξη, όπως την αναπαραγωγή, την μετάλλαξη, τον ανασυνδυασμό και την επιλογή. Οι υποψήφιες λύσεις για το πρόβλημα βελτιστοποίησης που θέλουμε να λύσουμε, διαδραματίζουν το ρόλο των

χρωμοσωμάτων σε έναν πληθυσμό, με τη συνάρτηση καταλληλότητας να καθορίζει το περιβάλλον μέσα στο οποίο αυτές ζούν. Ακολουθώς λαμβάνει χώρα εξέλιξη του πληθυσμού, μετά την επανειλημμένη εφαρμογή των πιο πάνω μηχανισμών. Οι εξελικτικοί αλγόριθμοι συχνά αποδίδουν καλά στην βελτιστοποίηση λύσεων, σε όλα τα είδη των προβλημάτων, διότι δεν κάνουν καμία υπόθεση για την υποκείμενη καταλληλότητα του τοπίου. Εκτός από τη χρήση τους ως μαθηματικοί βελτιστοποιητές, οι εξελικτικοί αλγόριθμοι έχουν επίσης χρησιμοποιηθεί και ως πειραματικό πλαίσιο για την επικύρωση θεωριών, σχετικά με τη βιολογική εξέλιξη και τη φυσική επιλογή.

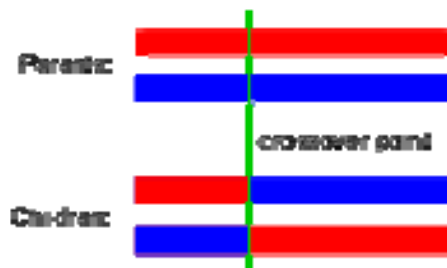
3.4 Γενετικοί αλγόριθμοι

Οι γενετικοί αλγόριθμοι αποτελούν μια ξεχωριστή κατηγορία των εξελικτικών αλγορίθμων, οι οποίοι χρησιμοποιούν τους μηχανισμούς που αναφέραμε πιο πάνω. Μπορούν να υλοποιηθούν σε μια υπολογιστική προσομοίωση, στην οποία ένας πληθυσμός από αφαιρετικές αναπαραστάσεις υποψηφίων λύσεων για ένα πρόβλημα βελτιστοποίησης, (οι οποίες ονομάζονται χρωμοσώματα), εξελίσσεται σε καλύτερες λύσεις. Παραδοσιακά, οι λύσεις αναπαρίστανται στο δυαδικό σύστημα ως συμβολοσειρές με 0s και 1s, αλλά και άλλες κωδικοποιήσεις είναι επίσης δυνατές. Η εξέλιξη ξεκινά συνήθως από έναν πληθυσμό με τυχαία άτομα και συμβαίνει σε γενιές. Σε κάθε γενιά, η καταλληλότητα του κάθε ατόμου του πληθυσμού αξιολογείται. Πολλαπλά άτομα επιλέγονται στοχαστικά από το σημερινό πληθυσμό (με βάση την καταλληλότητά τους), και τροποποιούνται για να σχηματίσουν ένα νέο πληθυσμό. Ο νέος πληθυσμός χρησιμοποιείται στην επόμενη επανάληψη του αλγορίθμου. Συνήθως, ο αλγόριθμος τερματίζεται όταν είτε ένας μέγιστος αριθμός γενεών έχει παραχθεί, ή ένα ικανοποιητικό επίπεδο καταλληλότητας έχει επιτευχθεί για τον πληθυσμό. Αν ο αλγόριθμος τερματιστεί λόγω του μέγιστου αριθμού γενεών, μπορεί να μην έχει επιτευχθεί ικανοποιητική λύση. Ένας τυπικός γενετικός αλγόριθμος απαιτεί μια γενετική αναπαράσταση του συνόλου των λύσεων (συνήθως πίνακας bits) και μια συνάρτηση καταλληλότητας για την αξιολόγηση της ποιότητας των λύσεων. Στη συνέχεια προχωρεί στην αρχικοποίηση ενός πληθυσμού με τυχαίες λύσεις, οι οποίες καλύπτουν το σύνολο των δυνατών λύσεων. Το μέγεθος του πληθυσμού εξαρτάται συνήθως από τη φύση του προβλήματος. Κατά τη διάρκεια κάθε γενιάς, ένα ποσοστό του υφιστάμενου πληθυσμού επιλέγεται για την ανατροφή της νέας γενιάς. Κάποια

άτομα του πληθυσμού επιλέγονται με μια διαδικασία που βασίζεται στην αξιολόγηση της καταλληλότητας τους και στην οποία τα καταλληλότερα έχουν περισσότερες πιθανότητες να επιλεγούν. Το επόμενο βήμα είναι να παραχθεί ένας πληθυσμός δεύτερης γενιάς από τα επιλεγμένα άτομα, με τη χρήση των μηχανισμών της διασταύρωσης και της μετάλλαξης, οι οποίοι επεξηγούνται πιο κάτω:

Διασταύρωση:

Γιά κάθε χρωμόσωμα στον πληθυσμό παράγεται ένας τυχαίος αριθμός. Εάν αυτός ο αριθμός είναι μικρότερος από την πιθανότητα διασταύρωσης, την οποία ορίζουμε εμείς, τότε το χρωμόσωμα αυτό επιλέγεται για διασταύρωση. Εάν ο συνολικός αριθμός χρωμοσωμάτων που έχουν επιλεγθεί είναι μονός, τότε είτε επιλέγουμε ακόμα ένα χρωμόσωμα, ή αφήνουμε εκτός ένα, έτσι ώστε να έχουμε ζύγο αριθμό χρωμοσωμάτων προς διασταύρωση. Στη συνέχεια παίρνουμε ανά δύο τα χρωμοσώματα και επιλέγεται ένα κοινό σημείο και στα δύο. Όλα τα δεδομένα (bits) μετά από το συγκεκριμένο σημείο, ανταλλάσσονται μεταξύ των δύο γονέων. Το αποτέλεσμα είναι τα παιδιά, όπως φαίνεται πιο κάτω.



Σχήμα 3.2: Διασταύρωση

Μετάλλαξη:

Γιά κάθε bit κάθε χρωμοσώματος στον πληθυσμό παράγεται ένας τυχαίος αριθμός. Εάν αυτός ο αριθμός είναι μικρότερος από την πιθανότητα μετάλλαξης, την οποία ορίζουμε εμείς, τότε το bit αυτό επιλέγεται για μετάλλαξη. Σε περίπτωση που το bit είναι '0', γίνεται '1' και αντίστροφα, όπως φαίνεται στο Σχήμα 3.3.

1	1	1	0	1	0	0	0	0	0
1	1	0	0	1	0	0	1	0	0

Σχήμα 3.3: *Μετάλλαξη*

Βήματα γενετικού αλγορίθμου:

1. Αρχικοποίηση τυχαίου πληθυσμού
2. Αξιολόγηση της καταλληλότητας κάθε χρωμοσώματος στον πληθυσμό.
3. Επανάληψη των πιο κάτω μέχρι τον τερματισμό:
 - Επιλογή των καταλληλότερων ατόμων για αναπαραγωγή.
 - Παραγωγή νέας γενιάς με τη χρήση των μηχανισμών διασταύρωσης και της μετάλλαξης.
 - Αξιολόγηση της καταλληλότητας της νέας γενιάς.
 - Αντικατάσταση των λιγότερο ποιοτικών ατόμων του πληθυσμού, με καλύτερα άτομα από τη νέα γενιά.

3.5 Inputs Sensitivity Analysis

Το inputs sensitivity analysis είναι μια τεχνική η οποία μπορεί να αναγνωρίσει το ποσοστό συμβολής κάθε εισόδου του δικτύου, στην παραγόμενη έξοδο. Με τον τρόπο αυτό μπορεί να δώσει σημαντικές πληροφορίες, όσο αφορά την αναγνώριση των κυριότερων δεδομένων που καθορίζουν το αποτέλεσμα του δικτύου. Η τεχνική αυτή δουλεύει με βάση τα αθροίσματα των βαρών των εισόδων. Συγκεκριμένα η συμβολή κάθε εισόδου, εκφράζεται με τον αριθμό του αθροίσματος όλων των βαρών του προς το 1^ο κρυφό επίπεδο μόνο. Όσο μεγαλύτερος ο αριθμός αυτός τόσο μεγαλύτερη είναι και η συμβολή του στην έξοδο.

Κεφάλαιο 4

Μεθοδολογία

4.1 Σκοπός συστήματος	18
4.2 Dataset που χρησιμοποιήθηκε	18
4.3 Επεξεργασία δεδομένων	19
4.4 Πρώτη Μέθοδος: Νευρωνικό δίκτυο με αλγόριθμο Back – Propagation	20
4.5 Δεύτερη μέθοδος : Γενετικοί αλγόριθμοι	32
4.6 Τρίτη Μέθοδος: Inputs Sensitivity Analysis	37
4.7 Ενσωμάτωση animation στο σύστημα	42

4.1 Σκοπός συστήματος

Το σύστημα που υλοποίησα έχει ως σκοπό την επίτευξη της μέγιστης δυνατής επιτυχίας πρόβλεψης, όσον αφορά την αξιοπιστία ενός πελάτη. Συγκεκριμένα υλοποίησα 3 διαφορετικές μεθόδους, οι οποίες στοχεύουν τόσο στην επίτευξη καλού αποτελέσματος, όσο και στην εξεύρεση των σημαντικότερων παραμέτρων του δικτύου αλλά και των σημαντικότερων παραγόντων που επηρεάζουν την εκπαίδευση του δικτύου, σύμφωνα πάντα με το ανάλογο dataset που χρησιμοποιείται. Οι 3 αυτοί μέθοδοι επεξηγούνται πιο αναλυτικά στη συνέχεια.

4.2 Dataset που χρησιμοποιήθηκε

Το σύστημα μου χρησιμοποιεί για την διεξαγωγή έρευνας και εξαγωγή αποτελεσμάτων, το dataset που υπάρχει στην ιστοσελίδα του τμήματος στατιστικής του πανεπιστημίου του Μονάχου. Λόγω του γεγονότος ότι δεν είχα στη διάθεση μου κάποιο dataset από κάποια κυπριακή τράπεζα, κατάλληλο για τη διεξαγωγή της έρευνας μου, αναγκάστηκα να χρησιμοποιήσω το προαναφερόμενο, το οποίο περιέχει 1000 δείγματα πελατών. Κάθε δείγμα περιλαμβάνει 20 παράγοντες – χαρακτηριστικά, μαζί με την πραγματική

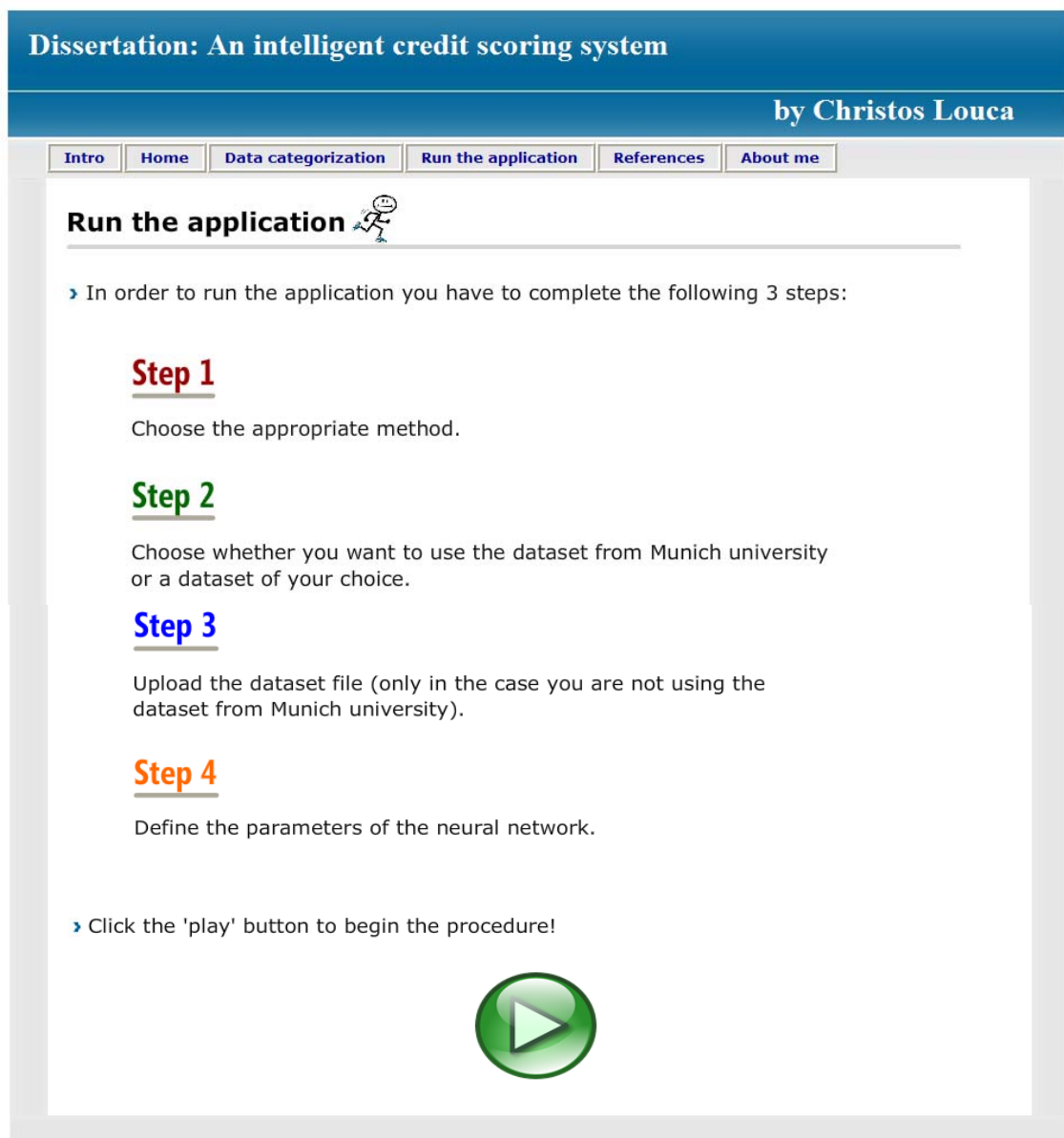
κατάσταση αξιοπιστίας του πελάτη ('1' για καλοπληρωτής και '0' για κακοπληρωτής). Τα χαρακτηριστικά αυτά είναι: κατάσταση τρέχοντος λογαριασμού, διάρκεια του λογαριασμού σε μήνες, πληρωμή προηγούμενων δανείων, σκοπός δανείου, πόσο δανείου σε ευρώ, αξία μετοχών και οικονομιών, χρόνος που εργάζεται για τον τωρινό εργοδότη, δόση υπό μορφή ποσοστού του διαθέσιμου εισοδήματος, οικογενειακή κατάσταση και φύλο, εγγυητές ή οφειλέτες, χρόνια ζωής στο τωρινό νοικοκυριό/σπίτι, περιουσιακά στοιχεία, ηλικία, επιπλέον ενεργά δάνεια, τύπος σπιτιού, επάγγελμα, αριθμός προηγούμενων δανείων στην συγκεκριμένη τράπεζα, τηλέφωνο, ξένος εργάτης και τέλος αριθμός ατόμων προς συντήρηση. Οι τιμές των παραγόντων αυτών δεν ήταν διαθέσιμες σε συνεχή μορφή, αλλά σε κατηγοριοποιημένη. Συγκεκριμένα μια συνεχής τιμή, ανάλογα με την κατηγορία που συσχετίζεται, παίρνει και την ανάλογη κατηγοριοποιημένη τιμή, συνήθως στο εύρος 0-10. Η νέα αυτή τιμή αναπαριστά την κατηγορία που ανήκει η συνεχής τιμή. Στο παράρτημα Α υπάρχει σχετικός πίνακας στον οποίο φαίνονται οι κατηγορίες κάθε γνωρίσματος και οι τιμές που παίρνουν, σύμφωνα με την ιστοσελίδα του πανεπιστημίου. Υπάρχουν επίσης γραφικές παραστάσεις που δείχνουν την κατανομή των δειγμάτων, σε σχέση με κάθε γνώρισμα ξεχωριστά.

4.3 Επεξεργασία δεδομένων

Για κάθε μέθοδο, τα δεδομένα επεξεργάστηκαν με μια ιδιαίτερη μέθοδο. Αρχικά χώρισα το αρχείο σε 2 κομμάτια. Το ένα περιείχε το σύνολο των αξιόπιστων πελατών και το άλλο αυτό των μη αξιόπιστων. Τα δύο αυτά σύνολα ακολούθως χωρίστηκαν σε 10 τμήματα. Το δε σύνολο των αξιόπιστων δειγμάτων σε 10 μέρη των 70, ενώ το σύνολο των μη αξιόπιστων σε 10 μέρη των 30 δειγμάτων. Στη συνέχεια τα αντίστοιχα τμήματα των δύο συνόλων, ενώθηκαν, με αποτέλεσμα να σχηματιστούν 10 υποσύνολα δεδομένων, με 100 δείγματα το καθένα. Κάθε νευρωνικό δίκτυο έτρεχε 10 φορές, έχοντας κάθε φορά ως δεδομένα εκπαίδευσης 9 από τις 10 εκατοντάδες και δεδομένα επαλήθευσης μια ξεχωριστή εκατοντάδα κάθε φορά. Η αποδοτικότητα του δικτύου, υπολογίζεται από το μέσο όρο των αποτελεσμάτων του, για τις 10 επαναλήψεις.

4.4 Πρώτη μέθοδος: Νευρωνικό δίκτυο με αλγόριθμο Back – Propagation

Στην πρώτη μου μέθοδο κατασκεύασα ένα νευρωνικό δίκτυο, το οποίο χρησιμοποιεί τον αλγόριθμο Back – propagation για την εκπαίδευση του. Το δίκτυο έχει τη δυνατότητα υποστήριξης δύο κρυφών επιπέδων, ενώ χρησιμοποιεί τον ρυθμό εκπαίδευσης και την ορμή κατά την επεξεργασία των δεδομένων του. Συγκεκριμένα ο χρήστης αρχικά εισέρχεται στη σελίδα ‘Run the application’, όπως φαίνεται στην Σχήμα 4.1 και πατά το κουμπί ‘Play’ για να ξεκινήσει τη διαδικασία.



Σχήμα 4.1: Σελίδα ‘Run the application’

Ακολουθώς μεταφέρεται στην επόμενη σελίδα (Σχήμα 4.2), όπου μπορεί να επιλέξει μια εκ των 3 επιλογών, όσο αφορά τη μέθοδο που θέλει να χρησιμοποιήσει. Σε περίπτωση που δεν επιλέξει τίποτα και θελήσει να μεταφερθεί στην επόμενη σελίδα, του εμφανίζεται ανάλογο μήνυμα (Σχήμα 4.3). Να σημειώσω ότι η μέθοδος που περιγράφω, την παρούσα στιγμή είναι η 1^η.

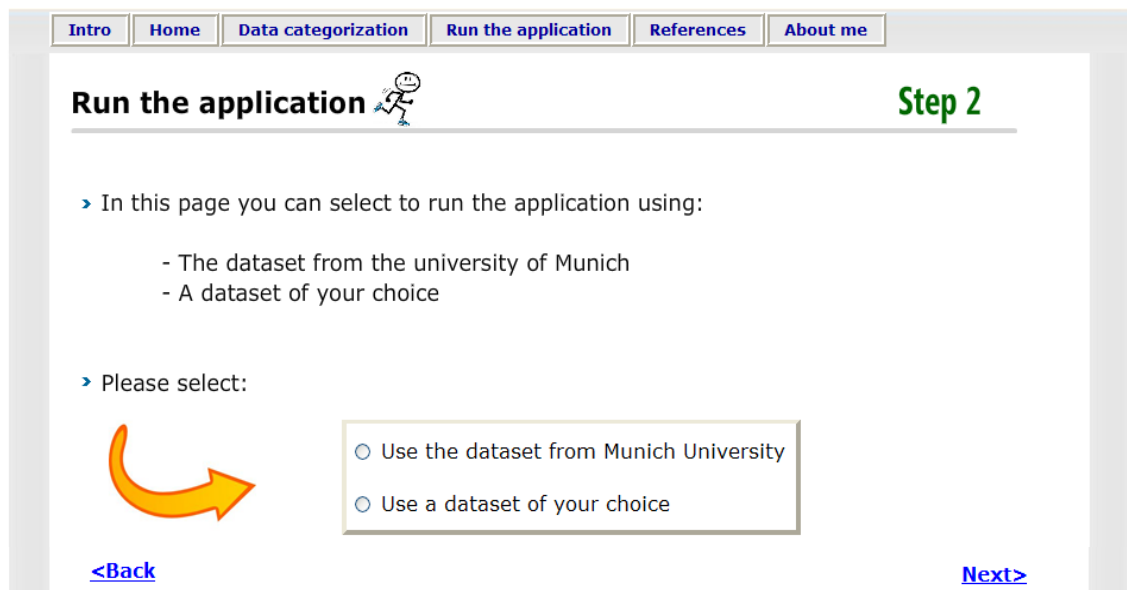
The screenshot shows a web application interface with a navigation bar at the top containing links: Intro, Home, Data categorization, Run the application (active), References, and About me. The main content area is titled 'Run the application' with a stick figure icon and 'Step 1' in red. Below the title, there is a paragraph: 'Here you can choose the preferred method to be used during the execution of the application.' followed by 'Please select:'. To the left of the selection box is a large red curved arrow pointing right. The selection box contains three radio buttons: 'Back-propagation algorithm' (selected), 'Genetic algorithm', and 'Inputs Sensitivity Analysis (ISA)'. At the bottom left is a '<Back' link and at the bottom right is a 'Next>' link.

Σχήμα 4.2: Σελίδα επιλογής μεθόδου

This screenshot is identical to the previous one, but it includes an error message at the bottom of the selection box: '* Please select something' in red text. The rest of the interface, including the navigation bar, title, instructions, arrow, and links, remains the same.

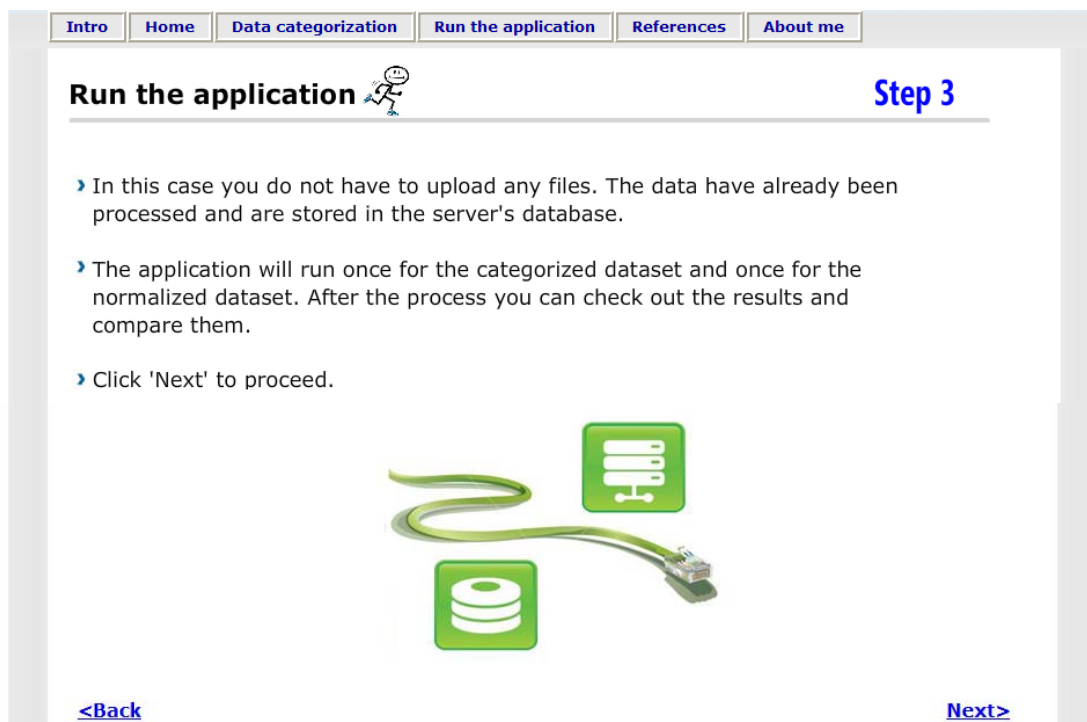
Σχήμα 4.3: Μήνυμα που εμφανίζεται σε περίπτωση λάθους

Στη συνέχεια ο χρήστης καλείται να επιλέξει το dataset, το οποίο θέλει να επεξεργαστεί από το σύστημα (Σχήμα 4.4). Μπορεί να επιλέξει να εργαστεί με τα δεδομένα του πανεπιστημίου του Μονάχου, τα οποία υπάρχουν στη βάση δεδομένων ή να ανεβάσει ο ίδιος τα δικά του αρχεία δεδομένων (excel ή text file). Υπάρχουν επίσης οι ανάλογοι ελέγχοι σε περίπτωση που δεν επιλεγθεί τίποτα από το χρήστη.



Σχήμα 4.4: Σελίδα επιλογής δεδομένων

Σε περίπτωση που ο χρήστης επιλέξει να εργαστεί με τα δεδομένα της βάσης, τότε αφού πατήσει το κουμπί 'Next', μεταφέρεται στην επόμενη σελίδα (Σχήμα 4.5), όπου ενημερώνεται για τον τρόπο με τον οποίο θα επεξεργαστούν τα δεδομένα. Συγκεκριμένα το πρόγραμμα για εκπαίδευση του δικτύου θα τρέξει δύο ξεχωριστές φορές, μια για τα δεδομένα σε κατηγοριοποιημένη μορφή και μια φορά για τα ίδια δεδομένα αλλά σε κανονικοποιημένη μορφή, δηλαδή στο εύρος τιμών 0 έως 1. Σκοπός είναι η σύγκριση των αποτελεσμάτων των δύο περιπτώσεων, έτσι ώστε να εξαχθούν συμπεράσματα ως προς την αποτελεσματικότητα του δικτύου, όσο και στον ρόλο που παίζει η μορφή των δεδομένων στην εκπαίδευση του δικτύου. Να σημειώσω ότι η κανονικοποίηση των τιμών έγινε με βάση την εξίσωση 4.1, η οποία εφαρμόστηκε ξεχωριστά για κάθε χαρακτηριστικό.



Σχήμα 4.5: Σελίδα ενημέρωσης του χρήστη

$$\text{Normalized value} = \frac{\text{Attribute value} - \text{Minimum attribute value}}{\text{Maximum attribute value} - \text{Minimum attribute value}}$$

Εξίσωση 4.1: Κανονικοποίηση δεδομένων

Ακολούθως ο χρήστης μεταφέρεται στην σελίδα όπου θα πρέπει να εισάγει τις παραμέτρους του δικτύου που επιθυμεί (Σχήμα 4.6). Εδώ ο αριθμός των εισόδων και εξόδων είναι προκαθορισμένος, αφού τα δεδομένα που χρησιμοποιούνται βρίσκονται στη βάση δεδομένων και είναι στατικά. Λόγω αυτού του γεγονότος, τα ανάλογα πεδία στη σελίδα δεν μπορούν να τροποποιηθούν από το χρήστη και περιέχουν σταθερές τιμές. Για την ενημέρωση του χρήστη υπάρχει ανάλογο μήνυμα στην αρχή της σελίδας. Σε περίπτωση που ο χρήστης κάνει οποιοδήποτε λάθος, είτε αυτό αφορά το εύρος τιμών ορισμένων πεδίων, είτε αφορά κενές τιμές, εμφανίζονται τα ανάλογα μηνύματα για τη σωστή εισαγωγή των παραμέτρων (Σχήμα 4.7). Οι υπόλοιπες παράμετροι αφορούν αριθμό νευρώνων στα δύο κρυφά επίπεδα, ρυθμό μάθησης, ορμή και επαναλήψεις.

IntroHomeData categorizationRun the applicationReferencesAbout me

Run the application

Step 4

> Customize the application parameters and click 'Run'.

! In this case the number of inputs and outputs is predefined and cannot be edited.

Number of inputs: 20
 Number of neurons in the first hidden layer:
 Number of neurons in the second hidden layer:
 Number of outputs: 1
 Learning rate:
 Momentum factor:
 Iterations:

[<Back](#)
[Run>>](#)

Σχήμα 4.6: Σελίδα παραμέτρων

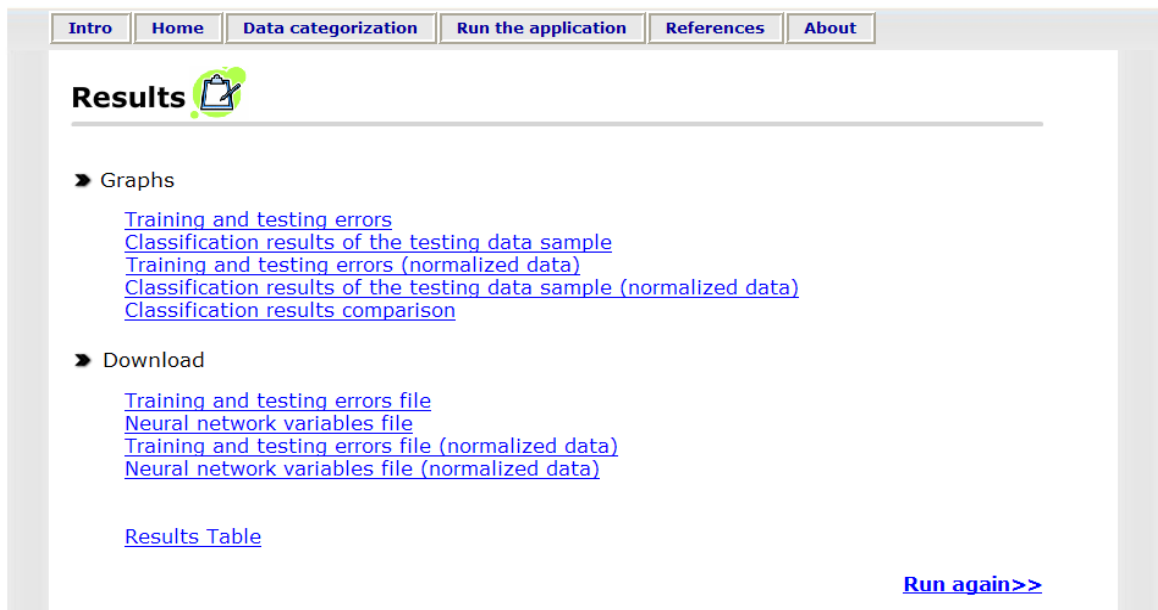
Number of inputs: 20
 Number of neurons in the first hidden layer: * Incomplete
 Number of neurons in the second hidden layer: 4, * Must be integer >= 0
 Number of outputs: 1
 Learning rate: 7 * Must be real in the range (0,1) e.g. 0,5
 Momentum factor: * Incomplete
 Iterations: * Incomplete

Σχήμα 4.7: Μηνύματα σε περίπτωση λαθών

Μόλις ο χρήστης πατήσει το κουμπί 'Run', αρχίζει η επεξεργασία των δεδομένων. Ανάλογα με τις τιμές των παραμέτρων ο χρόνος επεξεργασίας κάθε φορά είναι διαφορετικός. Με την ολοκλήρωση της επεξεργασίας ο χρήστης μεταφέρεται αυτόματα

24

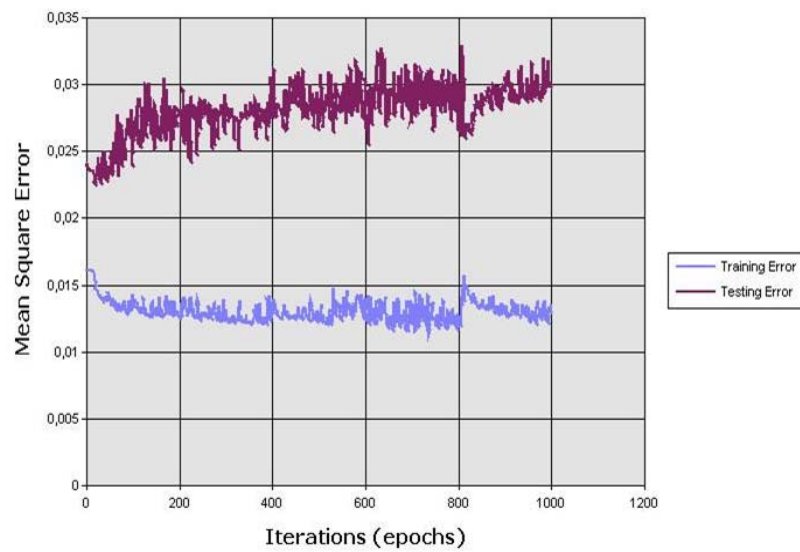
στη σελίδα των αποτελεσμάτων, όπου μπορεί να επιλέξει κάποιο σύνδεσμο που επιθυμεί (Σχήμα 4.8).



Σχήμα 4.8: Σελίδα αποτελεσμάτων 1^{ης} μεθόδου

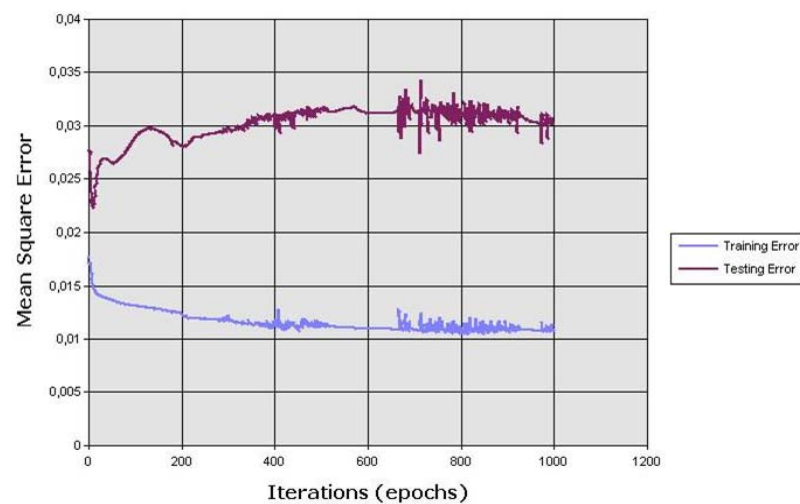
Οι επιλογές αυτές περιλαμβάνουν:

1. Τις γραφικές παραστάσεις των σφαλμάτων εκπαίδευσης και επαλήθευσης τόσο για τα κατηγοριοποιημένα, όσο και τα κανονικοποιημένα δεδομένα (Σχήμα 4.9 και Σχήμα 4.10).
2. Την αποτελεσματικότητα του δικτύου, μετά το τέλος της εκπαίδευσης, στο σύνολο των δεδομένων επαλήθευσης και για τις δύο ομάδες δεδομένων. Συγκεκριμένα τα δεδομένα επαλήθευσης περιέχουν 400 δείγματα από το αρχικό dataset (Σχήμα 4.11). Παρουσιάζονται ποσοστά επιτυχίας για τους καλοπληρωτές, κακοπληρωτές και μέσου όρου.
3. Τη σύγκριση της αποτελεσματικότητας των δύο τύπων δεδομένων (Σχήμα 4.12).
4. Το κατέβασμα αρχείων που περιέχουν αναλυτικά τα σφάλματα και τις τιμές των τελικών μεταβλητών (βαρών) του δικτύου (Σχήμα 4.13 και Σχήμα 4.14 αντίστοιχα).
5. Τον πίνακα με την αποτελεσματικότητα του δικτύου, για κάθε μια από τις 10 εκπαιδεύσεις του, τόσο για τα δεδομένα εκπαίδευσης όσο και για τα δεδομένα επαλήθευσης.



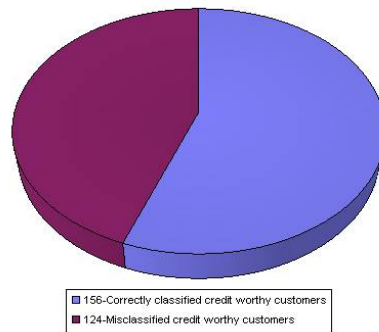
[<Back](#)

Σχήμα 4.9: Γραφική παράσταση σφαλμάτων κατηγοριοποιημένων δεδομένων

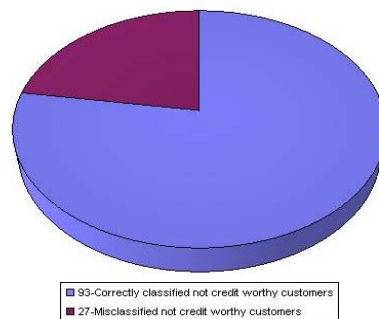


[<Back](#)

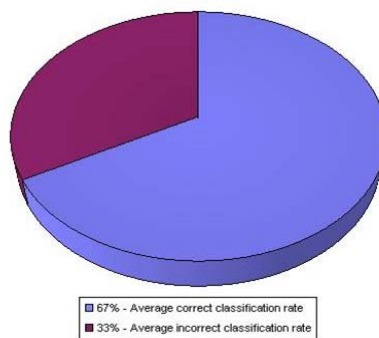
Σχήμα 4.10: Γραφική παράσταση σφαλμάτων κανονικοποιημένων δεδομένων



Total number of credit worthy customers:	280
Correctly classified:	156 (55,71%)
Incorrectly classified:	124 (44,29%)



Total number of not credit worthy customers:	120
Correctly classified:	93 (77,5%)
Incorrectly classified:	27 (22,5%)

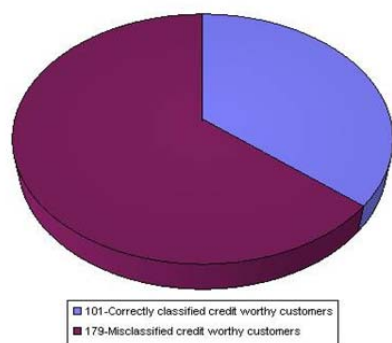


Average correct classification rate:	67%
Average incorrect classification rate:	33%

[<Back](#)

Σχήμα 4.11: Αποτελέσματα πρόβλεψης

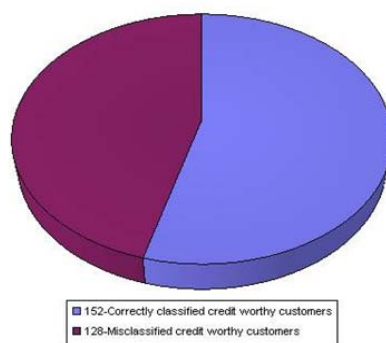
Not normalized data



Total number of credit worthy customers:	280
Correctly classified:	101 (36,07%)
Incorrectly classified:	179 (63,93%)

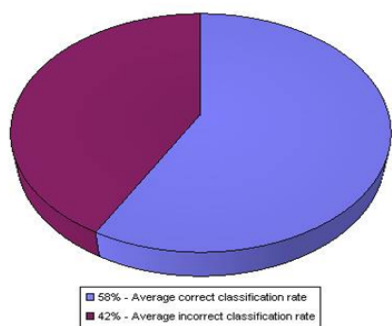
VS

Normalized data



Total number of credit worthy customers:	280
Correctly classified:	152 (54,29%)
Incorrectly classified:	128 (45,71%)

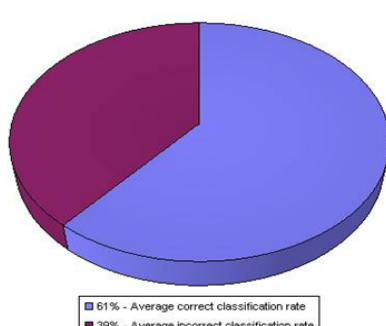
Not normalized data



Average correct classification rate:	58%
Average incorrect classification rate:	42%

VS

Normalized data



Average correct classification rate:	61%
Average incorrect classification rate:	39%

[<Back](#)

Σχήμα 4.12: Συγκριση αποτελεσμάτων για τις 2 ομάδες δεδομένων

Iteration Counter	Training Error	Testing Error
0	0,016239637725319	0,0252923441368511
1	0,0162347013646852	0,025313866066848
2	0,0162221581310893	0,0252017358550077
3	0,0161795703090993	0,0249614621888133
4	0,0161047955292409	0,0245719439925853
5	0,0160022184171612	0,0242156533670327
6	0,015864757167338	0,0238973953545551
7	0,0156892463396868	0,0237183847161915
8	0,0154664484151092	0,0239096136140603
9	0,0151870130296396	0,0242440944556607
10	0,0148616497010082	0,0246653880551228
11	0,0147244760598893	0,0252918370106086
12	0,0147136787252554	0,0247103943045805
13	0,0145707717177263	0,0254995811454138
14	0,0144875557748589	0,0259524527625767
15	0,0143418720033948	0,0269273224966992
16	0,0143850037943729	0,0269968424229651
17	0,0142386284995217	0,0272509853986571
18	0,0141378978890876	0,0262211744399253
19	0,0143818130989475	0,0266469814470438
20	0,0143108907912303	0,0260498412814224
21	0,0144801991483592	0,0262653053716704
22	0,0143678849327285	0,0263044188615627
23	0,0141830893766339	0,0264795124616486
24	0,0142534332817723	0,0279721845410991
25	0,0143395058712032	0,0267146315417521
26	0,014234012606202	0,0268425602106264
27	0,014212755441939	0,0272970964256761
28	0,0142872327452228	0,0272815155469214
29	0,0140465406755143	0,0270022653321753
30	0,0142050269158509	0,0263107174710896
31	0,0142163398952851	0,0267022437674842
32	0,0141958460672172	0,0267233489379223

Σχήμα 4.13: Στιγμιότυπο απο το αρχείο σφαλμάτων

Hidden Layer One:

Hidden Neuron: 0

Input: 4	Weight: 2,06079315729249
Input: 8	Weight: 1,65135068950259
Input: 4	Weight: 0,919140475853233
Input: 6	Weight: 0,158261095657498
Input: 7	Weight: 1,14270444460178
Input: 2	Weight: -0,0275318936726628
Input: 3	Weight: 3,58171582247358
Input: 4	Weight: -0,324853885308519
Input: 2	Weight: 2,45935714457814
Input: 1	Weight: -2,18485214129642

Σχήμα 4.14: Στιγμιότυπο από το αρχείο τελικών μεταβλητών

Training Set				Test Set	
Number	Actual Class	Good Credits	Bad Credits	Good Credits	Bad Credits
1	Good Credits	0	630	0	70
1	Bad Credits	270	0	30	0

Number	Actual Class	Good Credits	Bad Credits	Good Credits	Bad Credits
2	Good Credits	0	630	0	70
2	Bad Credits	270	0	30	0

Number	Actual Class	Good Credits	Bad Credits	Good Credits	Bad Credits
3	Good Credits	0	630	0	70
3	Bad Credits	270	0	30	0

Number	Actual Class	Good Credits	Bad Credits	Good Credits	Bad Credits
4	Good Credits	0	630	0	70
4	Bad Credits	270	0	30	0

Number	Actual Class	Good Credits	Bad Credits	Good Credits	Bad Credits
5	Good Credits	0	630	0	70
5	Bad Credits	270	0	30	0

Number	Actual Class	Good Credits	Bad Credits	Good Credits	Bad Credits
6	Good Credits	0	630	0	70
6	Bad Credits	270	0	30	0

Number	Actual Class	Good Credits	Bad Credits	Good Credits	Bad Credits
7	Good Credits	0	630	0	70
7	Bad Credits	270	0	30	0

Number	Actual Class	Good Credits	Bad Credits	Good Credits	Bad Credits
8	Good Credits	0	630	0	70
8	Bad Credits	270	0	30	0

Number	Actual Class	Good Credits	Bad Credits	Good Credits	Bad Credits
9	Good Credits	483	147	50	20
9	Bad Credits	187	83	21	9

Number	Actual Class	Good Credits	Bad Credits	Good Credits	Bad Credits
10	Good Credits	0	630	0	70
10	Bad Credits	270	0	30	0

Πίνακας 4.1: Πίνακας αποτελεσμάτων

Πιο πάνω φαίνεται ο πίνακας αποτελεσμάτων για κάθε μια από τις 10 εκπαιδεύσεις του δικτύου ξεχωριστά. Μπορούμε να δούμε πόσα δείγματα κατηγοριοποιήθηκαν σωστά, τόσο για τα δεδομένα εκπαίδευσης, όσο και για τα δεδομένα επαλήθευσης.

Τέλος σε περίπτωση που ο χρήστης επιλέξει να ανεβάσει κάποιο δικό του αρχείο, αντί να χρησιμοποιήσει αυτό της βάσης, μεταφέρεται από την σελίδα επιλογής δεδομένων (Σχήμα 4.4), στην σελίδα όπου θα ανεβάσει το αρχείο του (Σχήμα 4.15). Επίσης να αναφέρω ότι για διευκόλυνση του χρήστη, του ζητείται να ανεβάσει ένα μόνο αρχείο και να προσδιορίσει το ποσοστό που θέλει να επεξεργαστεί ως δεδομένα εκπαίδευσης. Το υπόλοιπο ποσοστό θα χρησιμοποιηθεί ως δεδομένα επαλήθευσης. Στη σελίδα αναγράφονται και οι κανόνες που πρέπει να τηρούν τα αρχεία έτσι ώστε να μην παρουσιάζονται προβλήματα στην επεξεργασία. Σε περιπτώσεις λάθους του χρήστη, όπως κενά πεδία, αριθμητικά σφάλματα ή ανέβασμα λάθους τύπου αρχείου, εμφανίζονται ανάλογα μηνύματα. Το σύστημα αναγνωρίζει αυτόματα τον τύπο αρχείου και ανάλογα επιτελεί την ανάλογη επεξεργασία. Μια άλλη μικρή διαφορά είναι ότι στη σελίδα των αποτελεσμάτων, ο χρήστης δεν θα μπορεί να δει συγκρίσεις αποτελεσμάτων, παρά μόνο τα αρχεία και τα αποτελέσματα των δικών του δεδομένων. Μετά από αυτή τη σελίδα τα βήματα που ακολουθούν (σελίδα παραμέτρων) παραμένουν τα ίδια, όπως περιγράφονται πιο πάνω.

Intro Home Data categorization Run the application References About me

Run the application Step 3

- › Upload the file containing the data and click 'Next'
- › Note that the file must obey the following rules:
 - › The file type should be .txt or .xls
 - › The first column must contain the output values.
 - › The values must be separated with a space character (.txt).
 - › Each line must contain only one pattern.
 - › No empty lines should exist in the files.
 - › The data must be categorized or normalized.
 - › It is recommended that there should be only one output.
 - › No empty lines should exist at the end of the files (.txt).
 - › It is recommended to express real numbers using ',' instead of '.'.
 - › It is recommended that the output values are either 0 for not credit worthy, either 1 for credit worthy.
 - › Each column must have a title (.xls).

Upload data file:

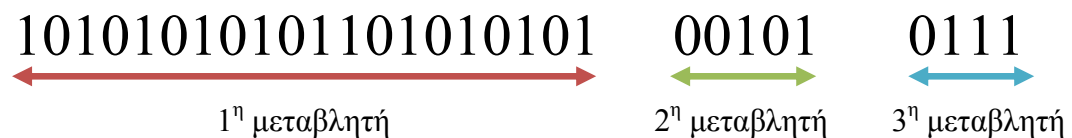
Give training sample percentage: %

[<Back](#) [Next>](#)

Σχήμα 4.15: Σελίδα για ανέβασμα αρχείου από χρήστη

4.5 Δεύτερη μέθοδος: Γενετικοί αλγόριθμοι

Στη δεύτερη μέθοδο χρησιμοποιήσα γενετικούς αλγόριθμους με σκοπό να προσπαθήσω να απομονώσω τις βέλτιστες τιμές παραμέτρων, του νευρωνικού μου δικτύου, οι οποίες βοηθούν στην καλύτερη εκπαίδευση του και στην εξαγωγή καλύτερων αποτελεσμάτων. Συγκεκριμένα ο πληθυσμός κάθε γενιάς αποτελείται από 20 χρωμοσώματα. Κάθε χρωμόσωμα αποτελείται από ένα αριθμό bits, τα οποία αναπαριστούν 3 μεταβλητές. Οι μεταβλητές αυτές είναι τα χαρακτηριστικά που θα χρησιμοποιηθούν στην εκπαίδευση, ο αριθμός των νευρώνων του 1^{ου} κρυφού επιπέδου και ο αριθμός των νευρώνων του 2^{ου} κρυφού επιπέδου. Για παράδειγμα σε περίπτωση που ο χρήστης χρησιμοποιήσει το έτοιμο dataset από τη βάση δεδομένων, το οποίο έχει 20 χαρακτηριστικά, το χρωμόσωμα θα αποτελείται από 29 bits όπως φαίνεται πιο κάτω:



Η 1^η μεταβλητή αποτελείται από τόσα bits, όσα και ο αριθμός των γνωρισμάτων κάθε φορά. Στο παράδειγμα μας 20. Εάν το αντίστοιχο bit κάθε γνωρίσματος είναι '1', τότε το γνώρισμα περιλαμβάνεται στην εκπαίδευση του δικτύου, αλλιώς εάν είναι '0', δεν λαμβάνεται υπόψη. Η 2^η και 3^η μεταβλητή αποτελούνται πάντα από 5 και 4 bits αντίστοιχα, αφού έθεσα ως όριο για τον αριθμό των κρυφών νευρώνων του 1^{ου} επιπέδου το 20 και για τον αριθμό των κρυφών νευρώνων του 2^{ου} επιπέδου το 10.

$$2^4 < 20 < 2^5 \longrightarrow \text{Άρα χρειάζονται 5 bits για την αναπαράσταση}$$

$$2^3 < 10 < 2^4 \longrightarrow \text{Άρα χρειάζονται 4 bits για την αναπαράσταση}$$

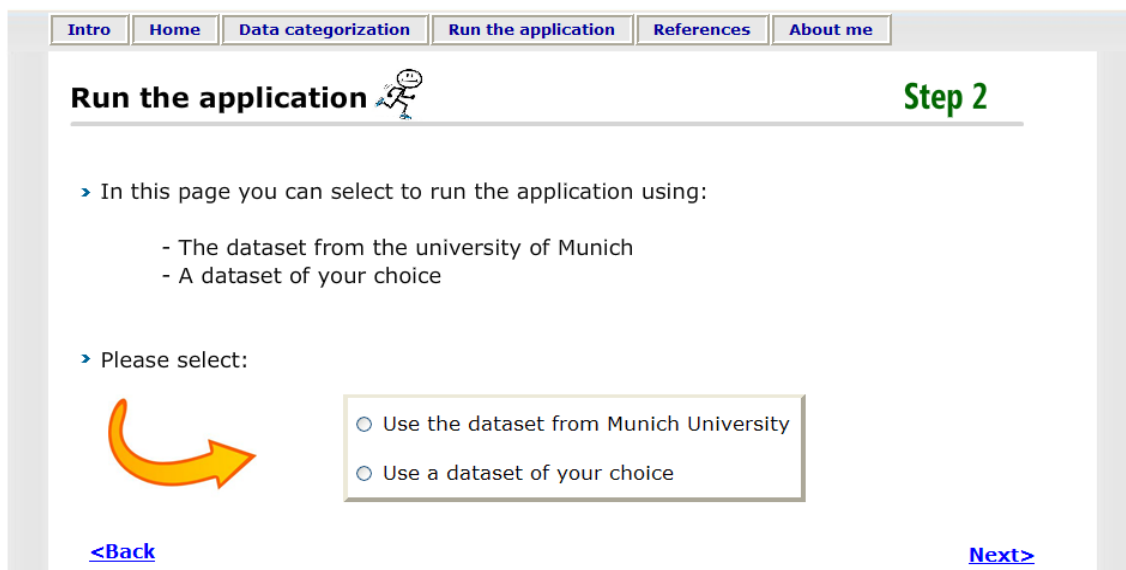
Όσον αφορά τον τρόπο με τον οποίο δουλεύει ο γενετικός αλγόριθμος, αρχικά δημιουργεί ένα τυχαίο πληθυσμό από 20 χρωμοσώματα. Η αξιολόγηση κάθε χρωμοσώματος γίνεται από την συνάρτηση καταλληλότητας, την οποία έθεσα ως το μέσο ποσοστό επιτυχίας που έχει το νευρωνικό δίκτυο της 1^{ης} μεθόδου, στα δεδομένα επαλήθευσης. Επομένως για την αξιολόγηση, το νευρωνικό δίκτυο θα πρέπει να εκπαιδευτεί για κάθε ξεχωριστό χρωμόσωμα, χρησιμοποιώντας κάθε φορά τις τιμές των παραμέτρων του χρωμοσώματος. Ακολούθως σχηματίζεται ένας νέος πληθυσμός, που βασίζεται στην επιλογή των καταλληλότερων χρωμοσωμάτων. Τέλος σε αυτόν το νέο πληθυσμό διενεργούνται οι μηχανισμοί της διασταύρωσης και της μετάλλαξης, όπως αυτοί περιγράφονται στο κεφάλαιο 3. Ο γενετικός αλγόριθμος επαναλαμβάνεται για 100 φορές. Με άλλα λόγια θα εκπαιδευτούν 2000 νευρωνικά δίκτυα. Πιο κάτω φαίνεται η σειρά βημάτων που απαιτείται για την επεξεργασία των δεδομένων με αυτή τη μέθοδο.

Αρχικά ο χρήστης επιλέγει τη δεύτερη μέθοδο στην παρακάτω εικόνα:

The screenshot shows a web application interface with a navigation bar at the top containing links: Intro, Home, Data categorization, Run the application (active), References, and About me. The main content area is titled 'Run the application' with a small stick figure icon and 'Step 1' in red. Below the title, it says 'Here you can choose the preferred method to be used during the execution of the application.' and 'Please select:'. There is a large red curved arrow pointing to a selection box. Inside the box, there are three radio button options: 'Back-propagation algorithm', 'Genetic algorithm' (which is selected), and 'Inputs Sensitivity Analysis (ISA)'. At the bottom left is a '<Back' link and at the bottom right is a 'Next>' link.

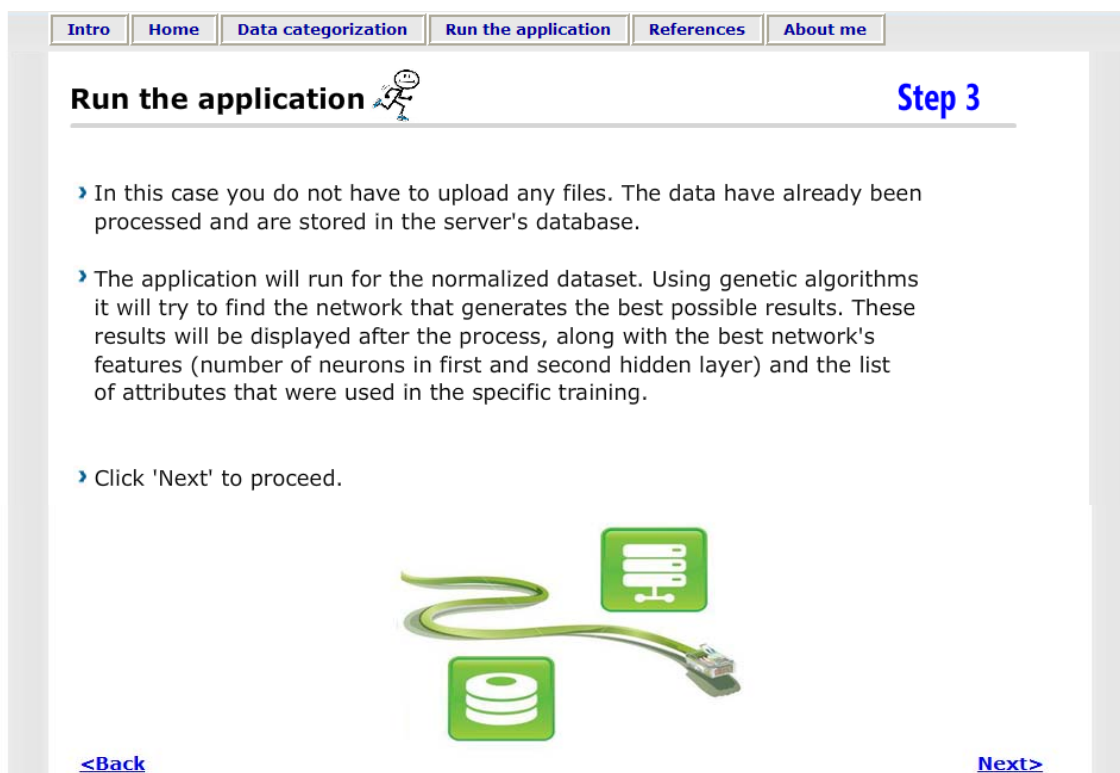
Σχήμα 4.16: Σελίδα επιλογής μεθόδου

Ακολούθως όπως και στην πρώτη μέθοδο, έτσι και εδώ, ο χρήστης καλείται να επιλέξει τα δεδομένα που θέλει να χρησιμοποιήσει (Σχήμα 4.17). Να σημειώσω ότι η συγκεκριμένη μέθοδος χρησιμοποιεί τα κανονικοποιημένα δεδομένα.



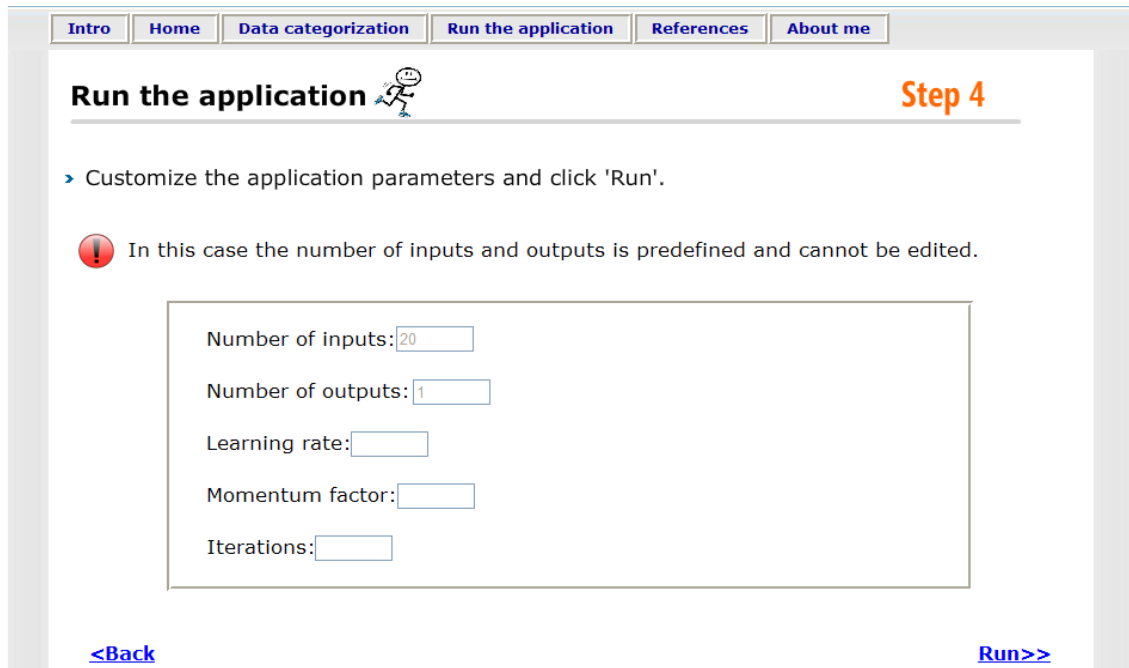
Σχήμα 4.17: Σελίδα επιλογής δεδομένων

Στη συνέχεια, στην περίπτωση που ο χρήστης επιλέξει να ανεβάσει δικό του αρχείο, μεταφέρεται στην ίδια σελίδα της 1^{ης} μεθόδου (Σχήμα 4.15), ενώ αν επιλέξει τα δεδομένα της βάσης, οδηγείται στην παρακάτω σελίδα:



Σχήμα 4.18: Σελίδα ενημέρωσης του χρήστη

Στη συνέχεια οδηγείται στη σελίδα των παραμέτρων (Σχήμα 4.19). Αυτή τη φορά οι παραμέτροι που πρέπει να εισάξει ο χρήστης είναι λιγότερες, αφού κάποιες από αυτές παράγονται από τον αλγόριθμο. Συγκεκριμένα ο χρήστης θα πρέπει να εισάξει αριθμό εισόδων, αριθμό εξόδων, ρυθμό μάθησης, ορμή και αριθμό επαναλήψεων. Οι τιμές αυτές θα ισχύουν για όλα τα νευρωνικά δίκτυα που πρόκειται να εκπαιδευτούν.



Intro Home Data categorization Run the application References About me

Run the application Step 4

> Customize the application parameters and click 'Run'.

❗ In this case the number of inputs and outputs is predefined and cannot be edited.

Number of inputs:

Number of outputs:

Learning rate:

Momentum factor:

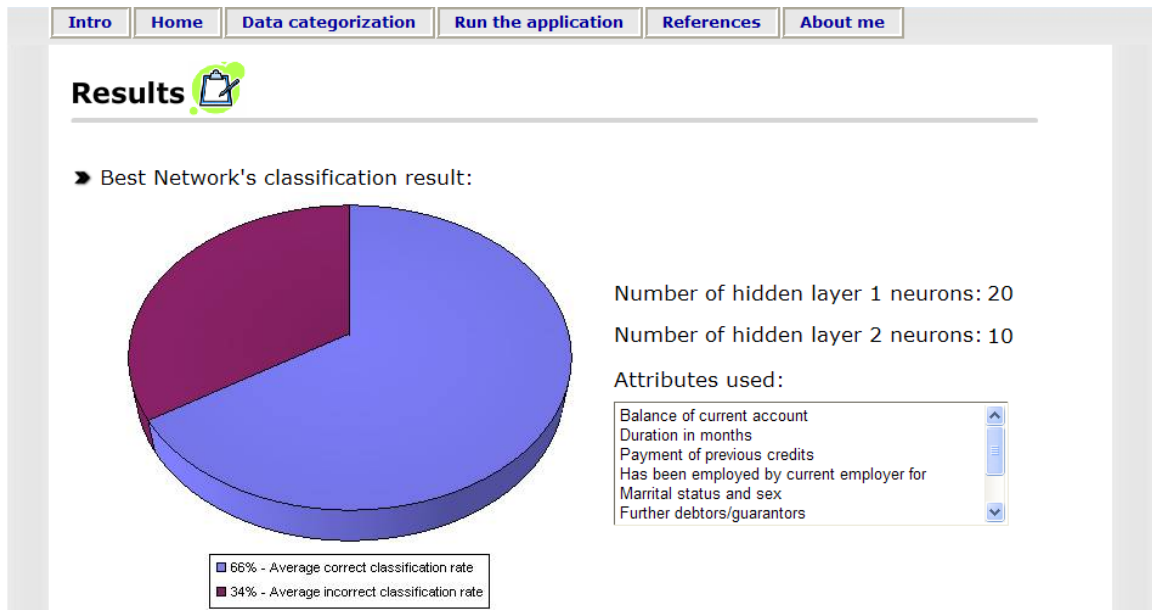
Iterations:

<Back Run>>

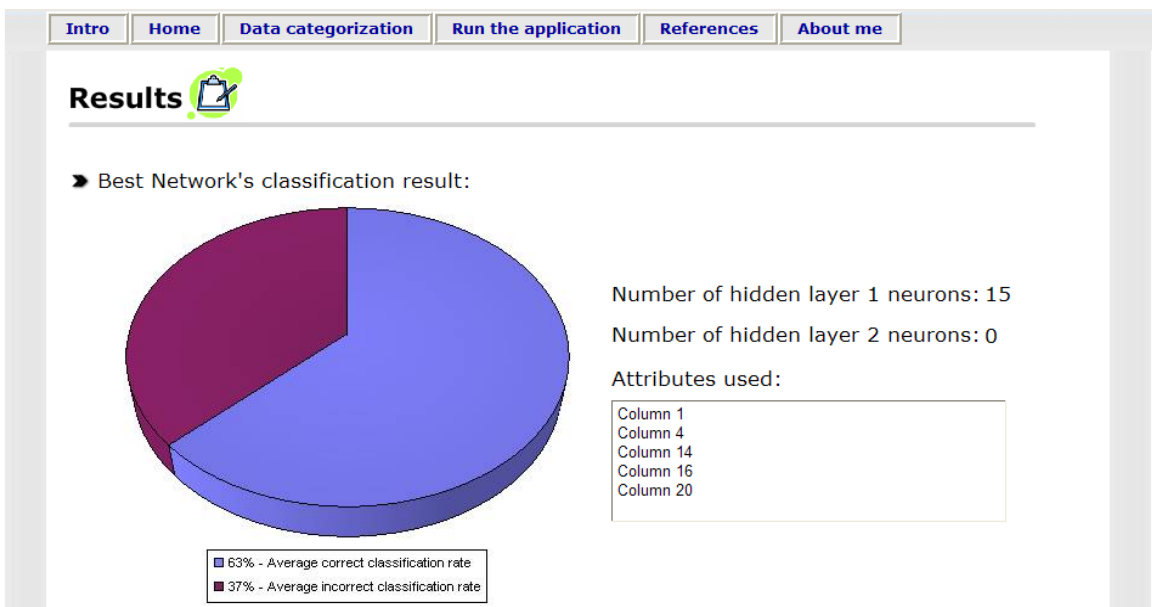
Σχήμα 4.19: Σελίδα εισαγωγής παραμέτρων

Τέλος αφού ολοκληρωθεί η επεξεργασία, αυτόματα παρουσιάζεται η σελίδα των αποτελεσμάτων (Σχήμα 4.20), η οποία περιέχει το μέσο ποσοστό επιτυχίας του καλύτερου δικτύου, καθώς και τους αριθμούς των νευρώνων για κάθε κρυφό επίπεδο. Επίσης σε μια λίστα περιέχει τα γνωρίσματα που χρησιμοποίησε το δίκτυο για να εκπαιδευτεί. Στην περίπτωση των δεδομένων της βάσης, αυτά αναγράφονται με το όνομα τους, αφού γνωρίζουμε σε ποιο γνώρισμα αντιστοιχεί κάθε στήλη. Στην περίπτωση δικού μας αρχείου, αναγράφεται απλά ο αριθμός της στήλης των γνωρισμάτων αυτών, με βάση τη σειρά που παρουσιάζονται στο αρχείο. Επομένως απλά ο χρήστης, το μόνο που θα έχει να κάνει, είναι να βρεί το γνώρισμα στο οποίο αντιστοιχεί η κάθε στήλη. Αυτό γίνεται αφού το σύστημα δεν μπορεί να γνωρίζει ποια χαρακτηριστικά χρησιμοποιούμε. Να σημειώσω ότι δεν υπάρχουν κάποια αρχεία για

κατέβασμα που αφορούν το δίκτυο, αφού σκοπός εδώ είναι η εύρεση των βέλτιστων παραμέτρων.



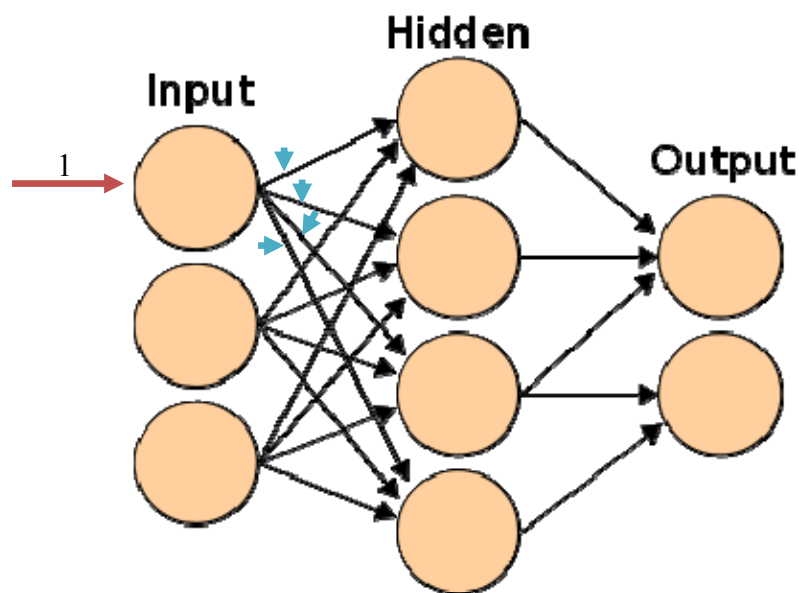
Σχήμα 4.20: Αποτελέσματα 2^{ης} μεθόδου (δεδομένα βάσης)



Σχήμα 4.20: Αποτελέσματα 2^{ης} μεθόδου (αρχείο χρήστη)

4.6 Μέθοδος 3^η: Inputs Sensitivity Analysis

Στην τρίτη και τελευταία μέθοδο χρησιμοποίησα την τεχνική “Inputs sensitivity analysis”. Η τεχνική αυτή προσπαθεί να βρεί το ποσοστό συμβολής κάθε χαρακτηριστικού, στην εκπαίδευση του δικτύου. Παρέχει επομένως σημαντική βοήθεια στην απομόνωση των σημαντικότερων παραγόντων, οι οποίοι επηρεάζουν περισσότερο την εκπαίδευση του δικτύου, όσο αφορά το θέμα του Credit scoring. Αρχικά εκπαιδεύονται 20 νευρωνικά δίκτυα. Ακολουθώντας με βάση το μέσο όρο επιτυχίας τους στο σύνολο των δεδομένων επαλήθευσης, επιλέγονται τα 4 πιο αποτελεσματικά. Στη συνέχεια για κάθε ένα από αυτά τα δίκτυα, υπολογίζεται το ποσοστό συμβολής κάθε ενός χαρακτηριστικού τους, που έλαβε μέρος στην εκπαίδευση. Αυτό γίνεται με την άρθρωση της απόλυτης τιμής των τελικών βαρών του 1^{ου} κρυφού επιπέδου μόνο, κάθε μιας εισόδου ξεχωριστά. Για παράδειγμα έστω ότι έχουμε το πιο κάτω νευρωνικό δίκτυο:



Σχήμα 4.21: Νευρωνικό δίκτυο

Για να βρούμε τον αριθμό ο οποίος θα εκφράζει το ποσοστό συμβολής της εισόδου 1, όπως φαίνεται στο σχήμα, θα πρέπει να αρθροίσουμε όλα τα βάρη τα οποία εκκινούν από τη συγκεκριμένη είσοδο και καταλήγουν σε νευρώνες του 1^{ου} κρυφού επιπέδου μόνο. Τα βάρη αυτά φαίνονται στο πιο πάνω σχήμα με γαλάζια βέλη. Για να

εκφράσουμε ως ποσοστό τον αριθμό αυτό, βρίσκουμε απλά το ολικό άθροισμα όλων αυτών των βαρών του 1^{ου} κρυφού επιπέδου και το υπολογίζουμε στη συνέχεια συγκρίνοντας το με τον αριθμό που βρήκαμε πιο πάνω. Εκτελούμε το ίδιο για κάθε είσοδο-χαρακτηριστικό. Με αυτό τον τρόπο μπορούμε να συγκρίνουμε τους σημαντικότερους παράγοντες που επηρέασαν κάθε δίκτυο ξεχωριστά και να εξάγουμε κάποιους γενικότερους κανόνες. Παρακάτω φαίνονται τα βήματα για εκτέλεση της πιο πάνω διαδικασίας.

Αρχικά επιλέγουμε τη μέθοδο, όπως φαίνεται στο Σχήμα 4.22:

Intro Home Data categorization Run the application References About me

Run the application

Step 1

> Here you can choose the preferred method to be used during the execution of the application.

Please select:



- ☐ Back-propagation algorithm
- ☐ Genetic algorithm
- ☒ Inputs Sensitivity Analysis (ISA)

<Back Next>

Σχήμα 4.22: Επιλογή μεθόδου

Intro Home Data categorization Run the application References About me

Run the application

Step 2

> In this page you can select to run the application using:

- The dataset from the university of Munich
- A dataset of your choice

> Please select:

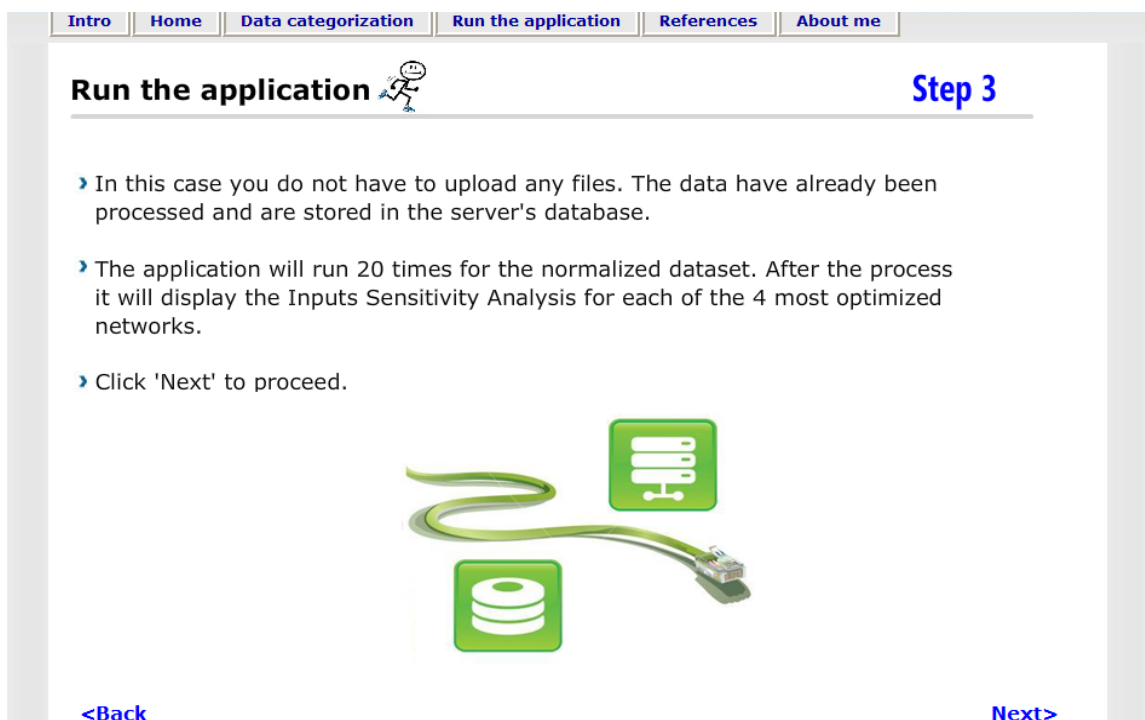


- ☐ Use the dataset from Munich University
- ☐ Use a dataset of your choice

<Back Next>

Σχήμα 4.23: Επιλογή δεδομένων

Ακολουθώντας όπως και στις προηγούμενες μεθόδους, έτσι και εδώ, ο χρήστης καλείται να επιλέξει τα δεδομένα που θέλει να χρησιμοποιήσει, όπως φαίνεται πιο πάνω στο Σχήμα 4.23. Να σημειώσω ότι η συγκεκριμένη μέθοδος χρησιμοποιεί και αυτή τα κανονικοποιημένα δεδομένα για αυξημένη αποτελεσματικότητα. Στη συνέχεια, στην περίπτωση που ο χρήστης επιλέξει να ανεβάσει δικό του αρχείο, μεταφέρεται στην ίδια σελίδα της 1^{ης} μεθόδου (Σχήμα 4.15), ενώ αν επιλέξει τα δεδομένα της βάσης, οδηγείται στην παρακάτω σελίδα:



Σχήμα 4.24: Σελίδα που παρουσιάζεται εάν ο χρήστης επιλέξει τη βάση δεδομένων

Στη συνέχεια πρέπει να καθορίσει τις παραμέτρους για το νευρωνικό δίκτυο, με τον ίδιο τρόπο όπως και στην 1η μέθοδο (Σχήμα 4.25). Σε περίπτωση λάθους εμφανίζονται τα ανάλογα μηνύματα.

[Intro](#)[Home](#)[Data categorization](#)[Run the application](#)[References](#)[About me](#)

Run the application

Step 4

Customize the application parameters and click 'Run'.

 In this case the number of inputs and outputs is predefined and cannot be edited.

Number of inputs:

Number of neurons in the first hidden layer:

Number of neurons in the second hidden layer:

Number of outputs:

Learning rate:

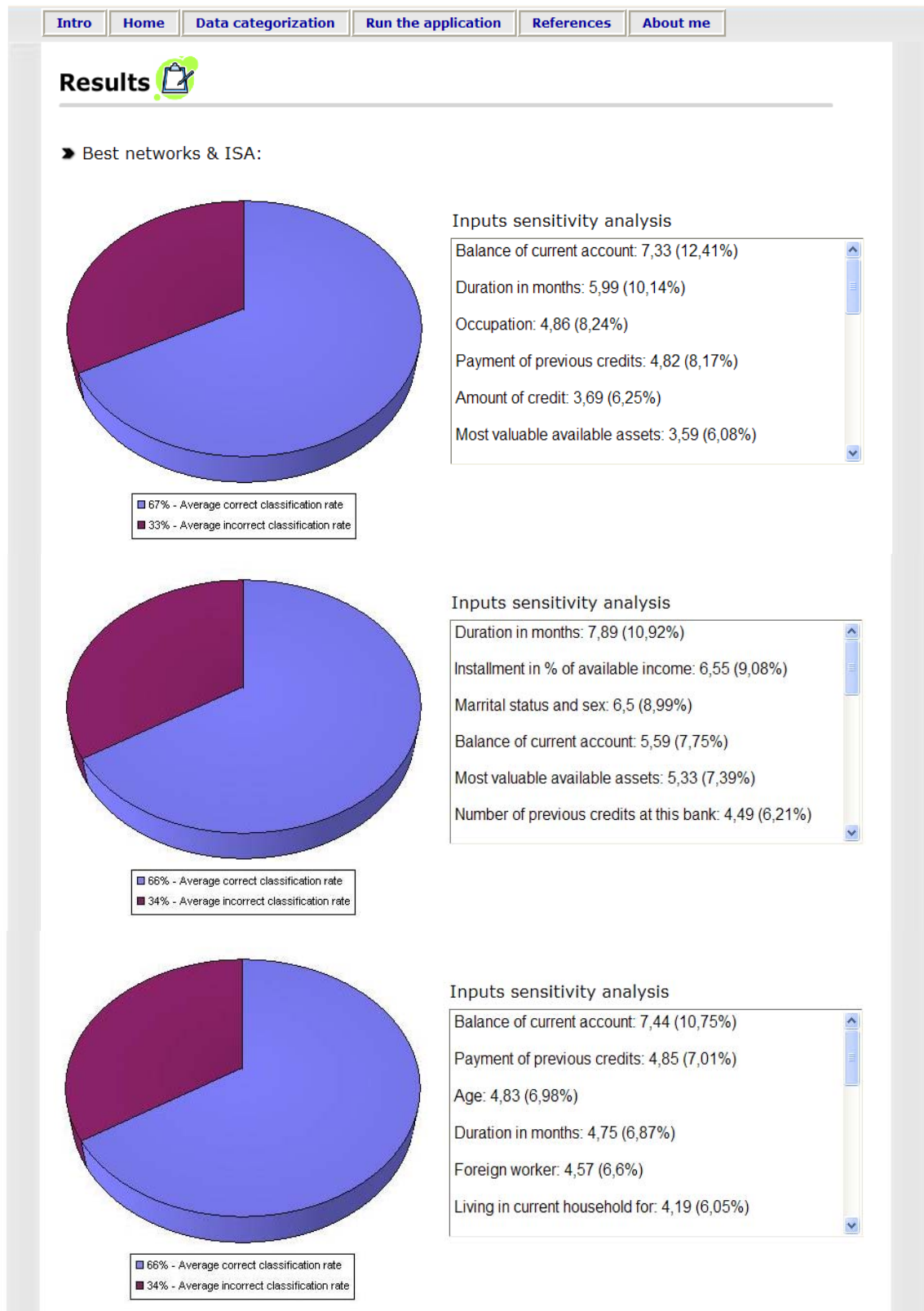
Momentum factor:

Iterations:

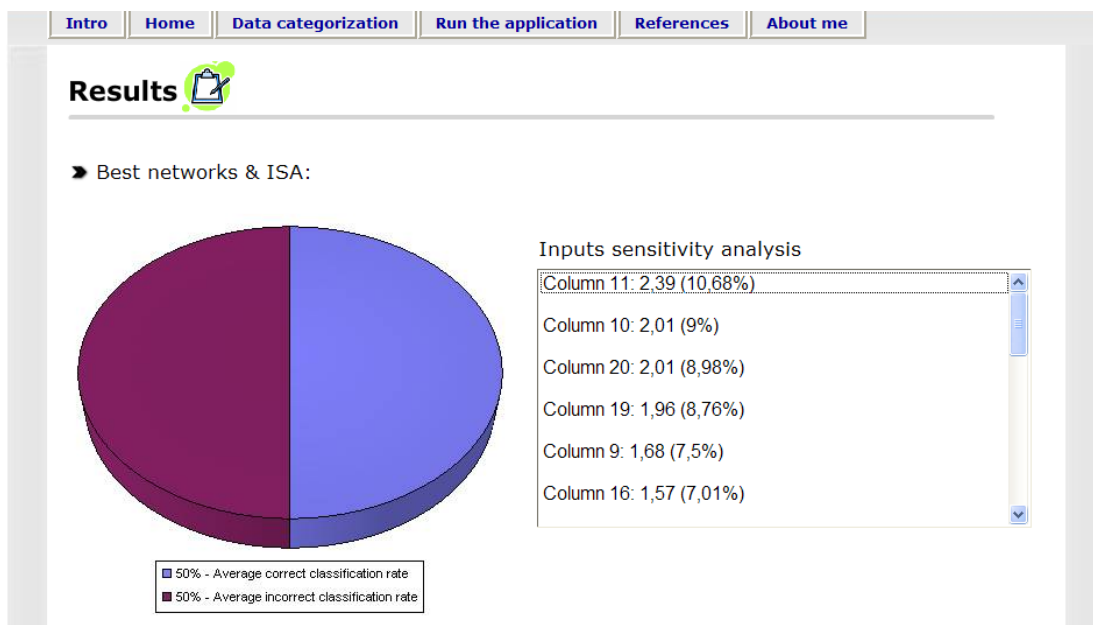
[<Back](#)[Run>>](#)

Σχήμα 4.25: Σελίδα εισαγωγής παραμέτρων

Μετά την επεξεργασία εμφανίζεται στο χρήστη η σελίδα με τα αποτελέσματα, η οποία περιέχει το μέσο ποσοστό επιτυχίας στο σύνολο των δεδομένων επαλήθευσης, υπό τη μορφή γραφικής παράστασης, για κάθε ένα από τα 4 πιο αποτελεσματικά δίκτυα. Επίσης για κάθε ένα από αυτά τα δίκτυα, υπάρχει και η λίστα με τα χαρακτηριστικά που χρησιμοποιήθηκαν μαζί με τα ποσοστά συμβολής τους στην εκπαίδευση. Εκτός από το ποσοστό, φαίνεται επίσης και ο αρχικός αριθμός που προκύπτει από την άρθρωση των βαρών. Να αναφέρω επίσης ότι σε περίπτωση που ο χρήστης χρησιμοποιήσει δικά του δεδομένα, οι λίστες παρουσιάζουν τα χαρακτηριστικά, ανάλογα με τον αριθμό της στήλης τους, στο αρχείο. Για παράδειγμα Column 2, αντιστοιχεί στη δεύτερη σε σειρά στήλη εισόδου, στο αρχείο. Οι μορφές των αποτελεσμάτων φαίνονται



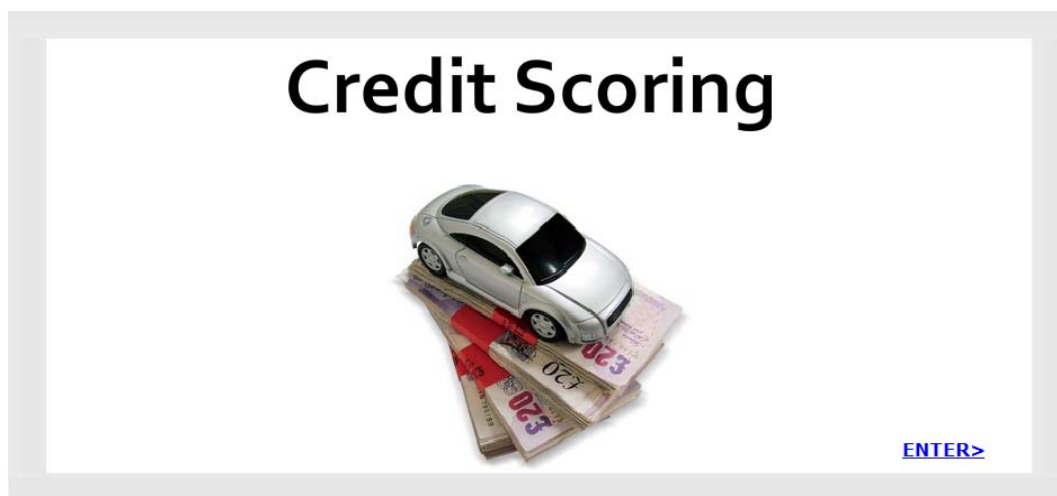
Σχήμα 4.26: Σελίδα αποτελεσμάτων 3^{ης} μεθόδου



Σχήμα 4.27: Μορφή αποτελεσμάτων για αρχείο του χρήστη

4.7 Ενσωμάτωση animation στο σύστημα

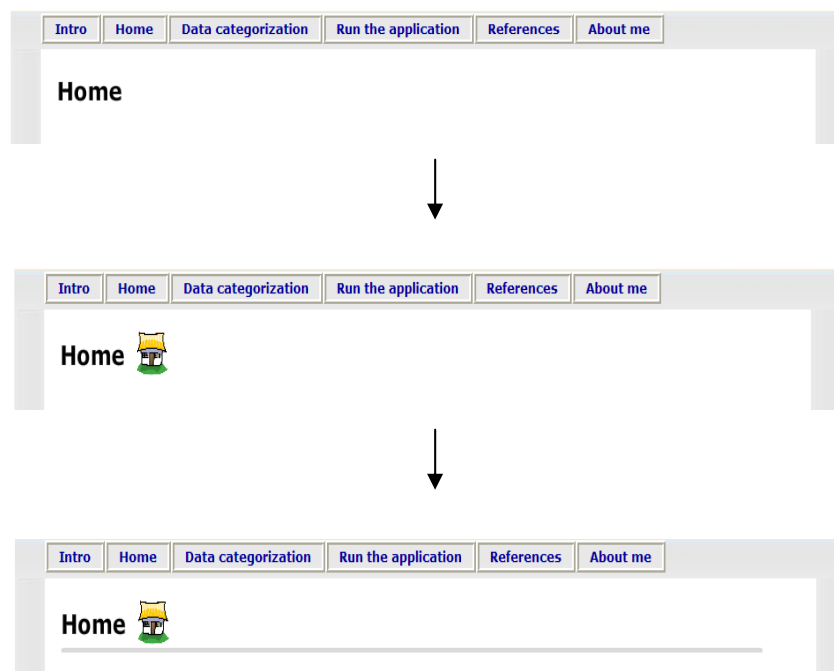
Σε κάποιες σελίδες του συστήματος υπάρχουν κινούμενες εικόνες και γράμματα (animation). Τα animation αυτά σχεδιάστηκαν με τη βοήθεια του Adobe Flash και αποσκοπούν στην προσθήκη ωραιότητας στο σύστημα. Μερικά από αυτά φαίνονται στις εικόνες πιο κάτω:



Σχήμα 4.28: Στιγμιότυπο από το animation εισαγωγής



Σχήμα 4.29: Στιγμιότυπο από το animation εισαγωγής



Σχήμα 4.30: Στιγμιότυπο από animation

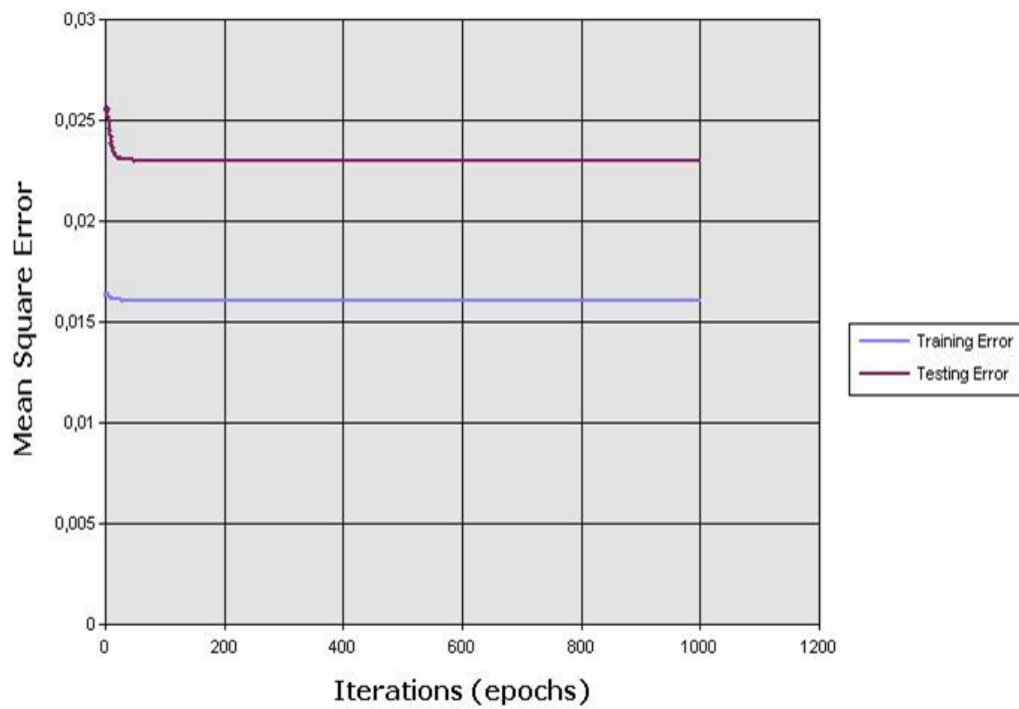
Κεφάλαιο 5

Αποτελέσματα

5.1 Αποτελέσματα πρώτης μεθόδου	44
5.2 Αποτελέσματα δεύτερης μεθόδου	54
5.3 Αποτελέσματα τρίτης μεθόδου	57

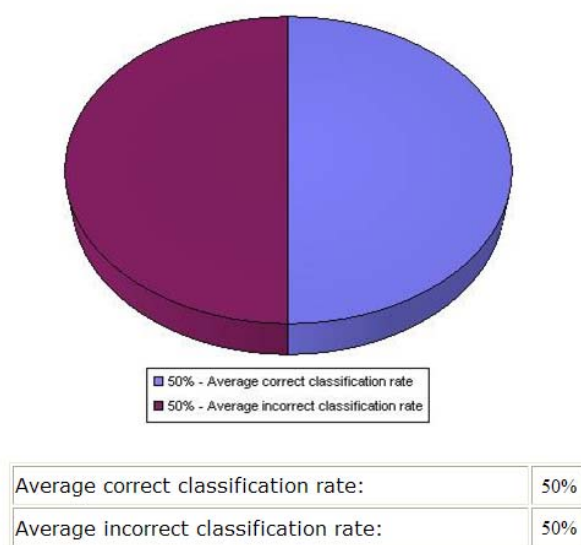
5.1 Αποτελέσματα πρώτης μεθόδου

Σκόπος της 1^{ης} μεθόδου είναι η επίτευξη ενός καλού ποσοστού επιτυχίας, στην πρόβλεψη της αξιοπιστίας των πελατών, για ένα σύνολο δεδομένων. Χρησιμοποιώντας το δείγμα δεδομένων του πανεπιστημίου του Μονάχου, εκτέλεσα αρκετές δοκιμές με διαφορετικές παράμετρους δικτύου κάθε φορά. Να σημειώσω ότι τα έτοιμα δεδομένα του Μονάχου, είναι σε κατηγοριοποιημένη μορφή, εκτός από κάποια γνωρίσματα και συγκεκριμένα η διάρκεια σε μήνες, η ηλικία και το ποσό του δανείου. Αρχικά εκτέλεσα τις δοκιμές με τα δεδομένα χωρίς να προχωρήσω σε οποιεσδήποτε αλλαγές στις τιμές τους. Χώρισα τα δεδομένα σε 2 ομάδες. Η μια ομάδα αποτελεί το σύνολο δεδομένων εκπαίδευσης το οποίο περιέχει 600 δείγματα (60% του συνολικού) και συγκεκριμένα 420 περιπτώσεις ‘καλοπληρωτών’ και 180 περιπτώσεις ‘κακοπληρωτών’. Η άλλη ομάδα αποτελεί το σύνολο δεδομένων επαλήθευσης και περιέχει 400 δείγματα, 280 περιπτώσεις ‘καλοπληρωτών’ και 120 περιπτώσεις ‘κακοπληρωτών’. Τα δεδομένα επαλήθευσης είναι αυτά τα οποία θα δοθούν στο δίκτυο μετά την εκπαίδευση του, έτσι ώστε να αξιολογηθεί η απόδοση του, ανάλογα με το ποσοστά επιτυχίας ή αποτυχίας πρόβλεψης που θα πετύχει στα δεδομένα αυτά. Μετά από αρκετές δοκιμές, κατέληξα στο συμπέρασμα ότι το δίκτυο δεν κατόρθωνε να εκπαιδευτεί σωστά. Πιθανή αιτία ήταν το γεγονός ότι τα δεδομένα δεν ήταν όλα κατηγοριοποιημένα στο σύνολο τους. Παρακάτω φαίνονται τα αποτελέσματα και οι γραφικές παραστάσεις των σφαλμάτων.



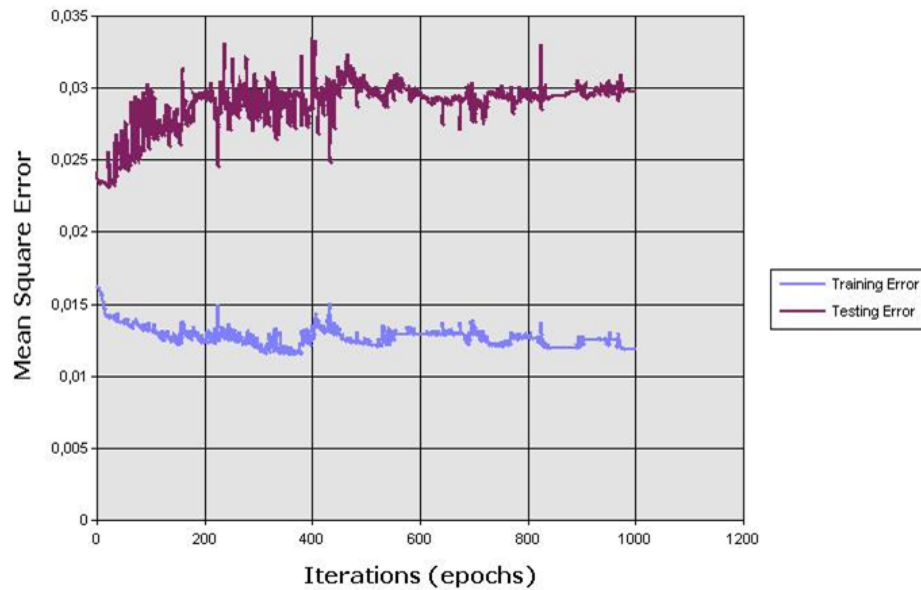
Σχήμα 5.1: Σφάλματα επαλήθευσης και εκπαίδευσης

Στο Σχήμα 5.1 πιο πάνω βλέπουμε ότι δεν υπάρχει σχεδόν καθόλου μείωση στο σφάλμα εκπαίδευσης, με αποτέλεσμα το δίκτυο να μην εκπαιδεύεται, όποιες και αν ήταν οι παράμετροι. Αυτό είχε ως αποτέλεσμα το μέσο ποσοστό επιτυχίας να ήταν πάντα 50%, το οποίο φυσικά δεν ήταν αληθινό, αφού η αξιοπιστία των δειγμάτων των πελατών επεξεργαζόταν από ανεκπαίδευτο δίκτυο και δεν είναι αντιπροσωπευτική.

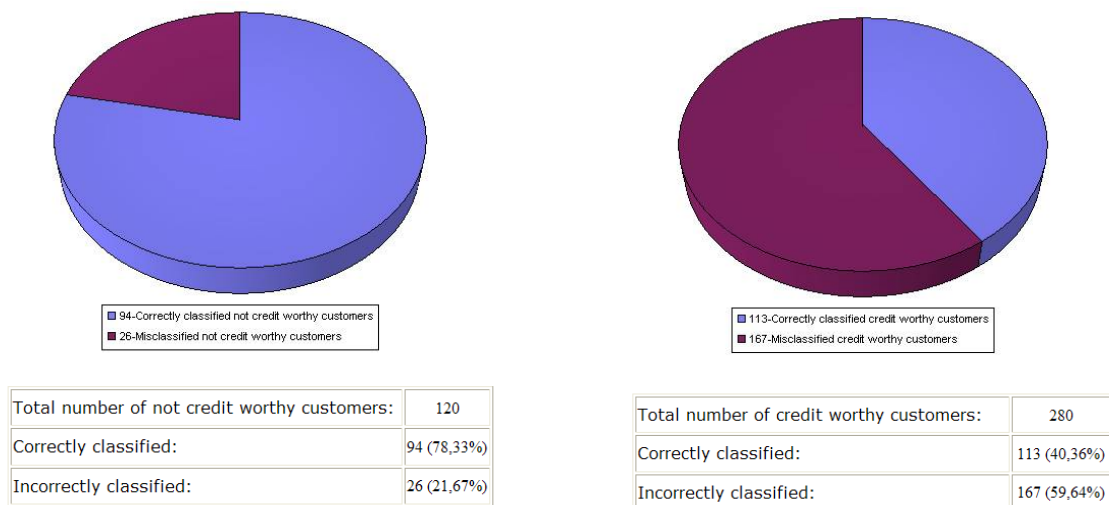


Σχήμα 5.2: Μέσο ποσοστό επιτυχίας ανεκπαίδευτου δικτύου

Στη συνέχεια αποφάσισα να κατηγοριοποιήσω όλα τα γνωρίσματα του συνόλου των δεδομένων, σε μια προσπάθεια μου να πετύχω την εκπαίδευση του δικτύου. Τα νέα αυτά δεδομένα τα αποθήκευσα στη βάση δεδομένων έτσι ώστε να τα χρησιμοποιήσω και σε μετέπειτα δοκιμές. Εκτελώντας αρκετά πειράματα παρατήρησα ότι είχα καταφέρει τον αρχικό μου στόχο, ο οποίος ήταν η εκπαίδευση του δικτύου. Κατάφερα να πετύχω αρκετά καλά αποτελέσματα, τα οποία έφταναν μέχρι και το 67%. Πιο κάτω μπορούμε να δούμε ότι υπάρχει κάποια μείωση στο σφάλμα επαλήθευσης, σε σχέση με τη περίπτωση ανεκπαίδευτου δικτύου του σχήματος 5.1.

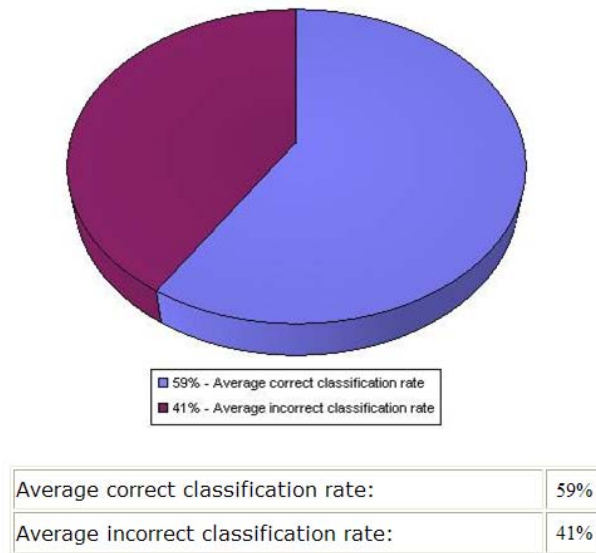


Σχήμα 5.3: Σφάλματα επαλήθευσης και εκπαίδευσης για τα κατηγοριοποιημένα δεδομένα



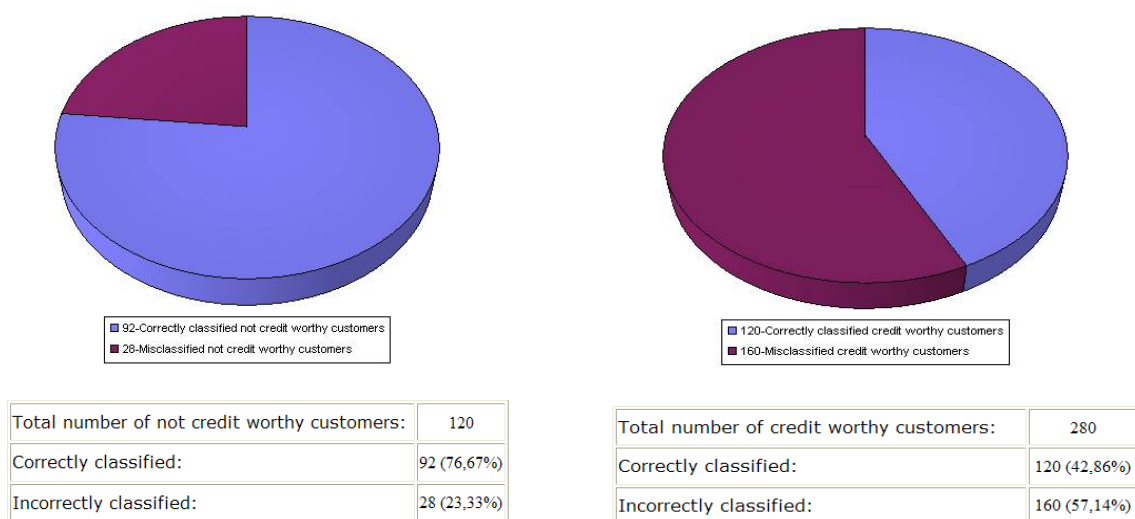
Σχήμα 5.4: 1^η σειρά αποτελεσμάτων για τα κατηγοριοποιημένα δεδομένα

Όπως φαίνεται στο Σχήμα 5.4, το ποσοστό επιτυχίας για ‘κακοπληρωτές’ ήταν 78,33%, αρκετά μεγαλύτερο από αυτό των ‘καλοπληρωτών’ που ήταν μόνο 40,36%. Έτσι το μέσο ποσοστό επιτυχίας έφτασε μέχρι και το 59% όπως φαίνεται πιο κάτω:

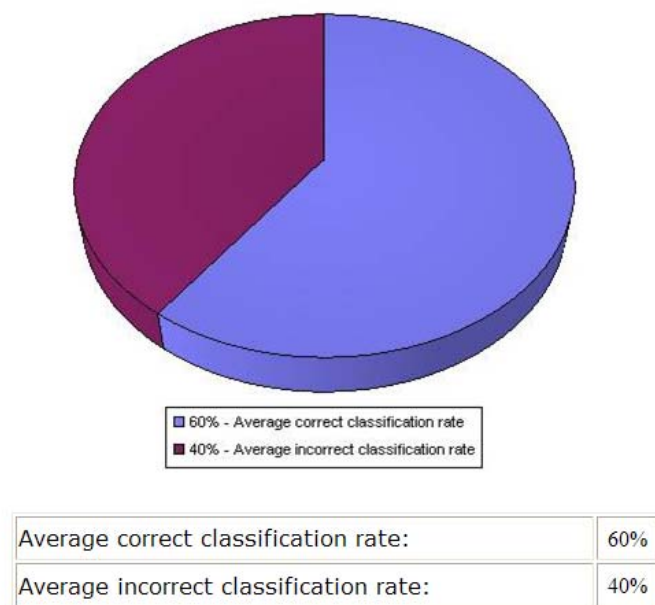


Σχήμα 5.5: Μέσο ποσοστό επιτυχίας 1^{ης} σειράς αποτελεσμάτων για τα κατηγοριοποιημένα δεδομένα

Σε άλλη σειρά αποτελεσμάτων όπως φαίνεται παρακάτω, το ποσοστό επιτυχίας για ‘κακοπληρωτές’ ήταν 76,67%, αρκετά μεγαλύτερο από αυτό των ‘καλοπληρωτών’ που ήταν μόνο 42,86%. Έτσι το μέσο ποσοστό επιτυχίας έφτασε μέχρι και το 60%.

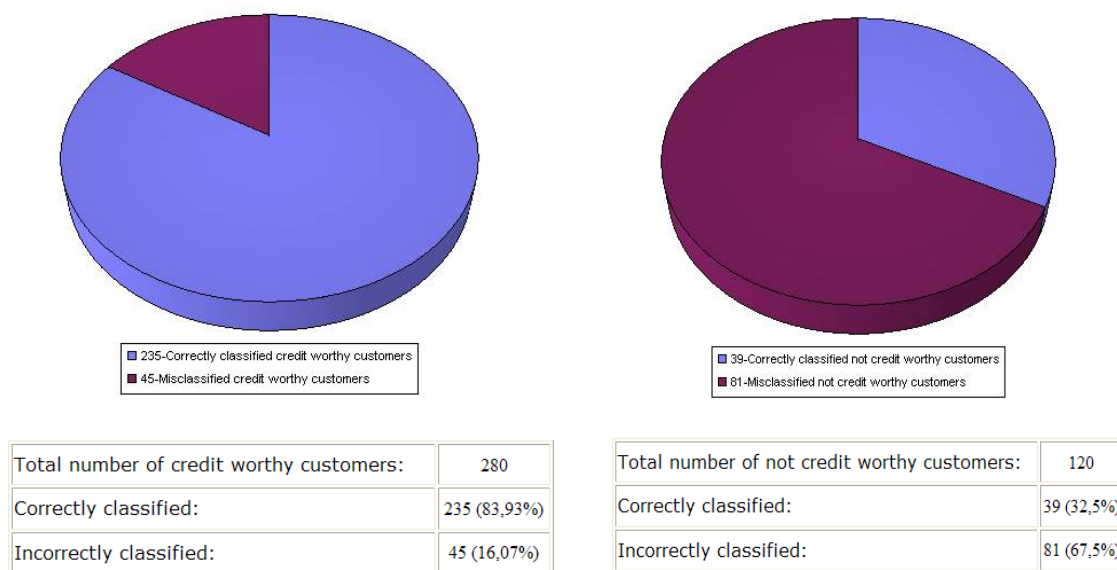


Σχήμα 5.6: 2^η σειρά αποτελεσμάτων για τα κατηγοριοποιημένα δεδομένα

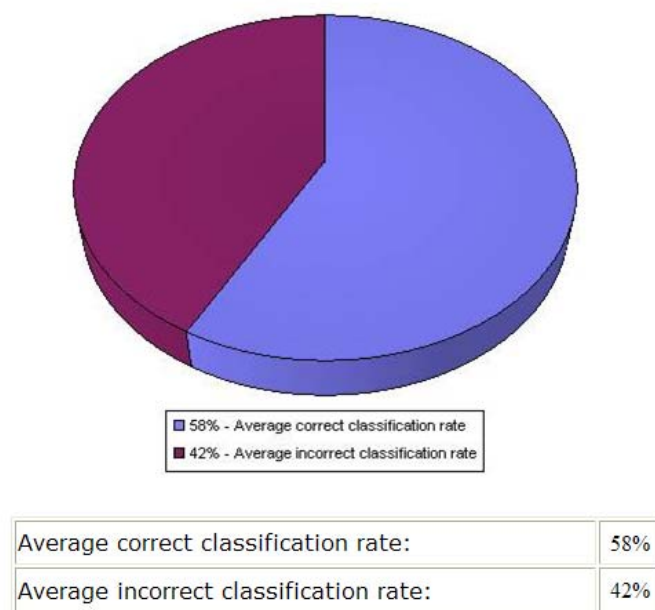


Σχήμα 5.7: Μέσο ποσοστό επιτυχίας 2^{ης} σειράς αποτελεσμάτων για τα κατηγοριοποιημένα δεδομένα

Σε άλλη σειρά αποτελεσμάτων όπως φαίνεται παρακάτω, το ποσοστό επιτυχίας για ‘καλοπληρωτές’ ήταν 83,93%, αρκετά μεγαλύτερο από αυτό των ‘κακοπληρωτών’ που ήταν μόνο 32,5%. Έτσι το μέσο ποσοστό επιτυχίας έφτασε μέχρι και το 58%.

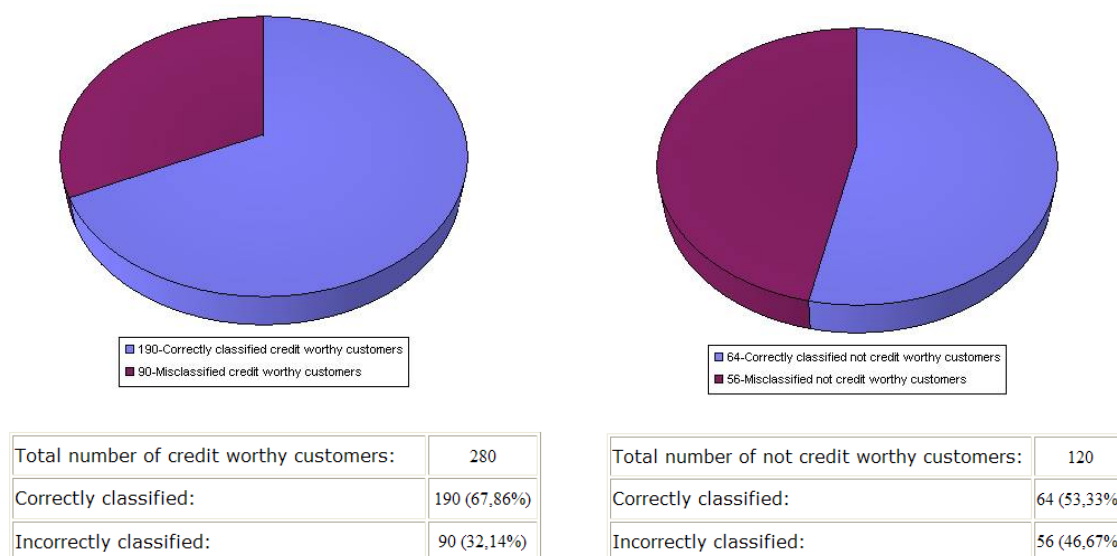


Σχήμα 5.8: 3^η σειρά αποτελεσμάτων για τα κατηγοριοποιημένα δεδομένα

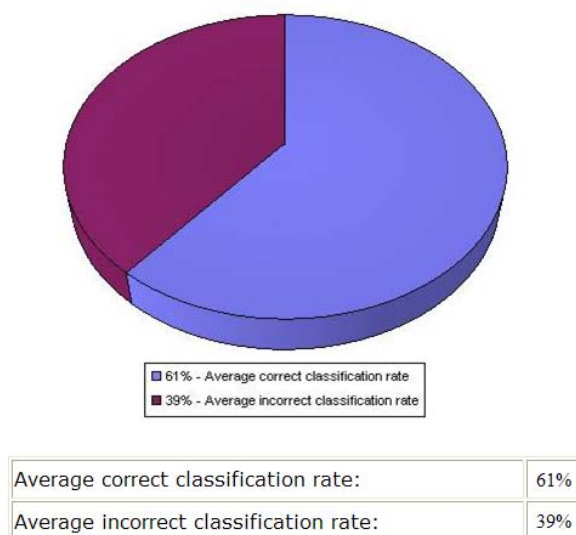


Σχήμα 5.9: Μέσο ποσοστό επιτυχίας 3^{ης} σειράς αποτελεσμάτων για τα κατηγοριοποιημένα δεδομένα

Στην 4^η σειρά αποτελεσμάτων όπως φαίνεται παρακάτω, το ποσοστό επιτυχίας για ‘καλοπληρωτές’ ήταν 67,86%, αρκετά μεγαλύτερο από αυτό των ‘κακοπληρωτών’ που ήταν μόνο 53,33%. Έτσι το μέσο ποσοστό επιτυχίας έφτασε μέχρι και το 61%.

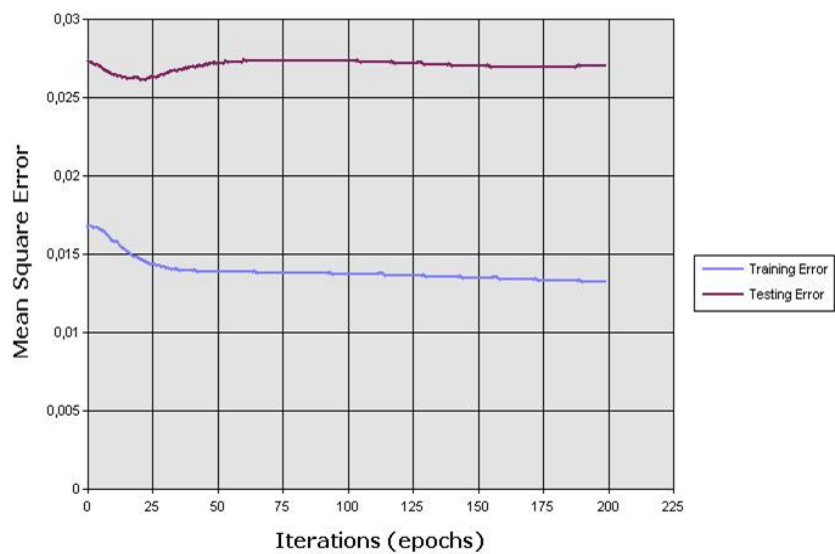


Σχήμα 5.10: 4^η σειρά αποτελεσμάτων για τα κατηγοριοποιημένα δεδομένα



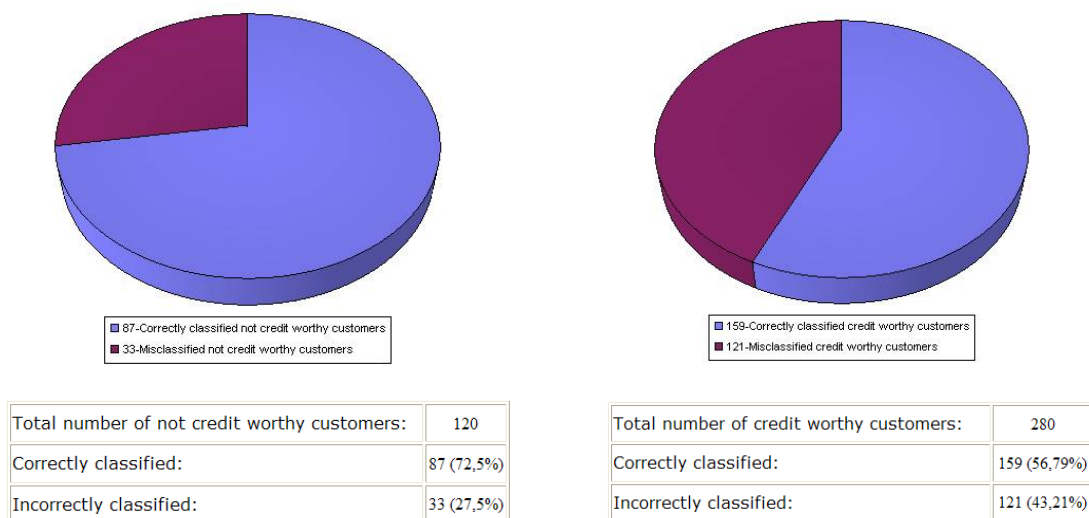
Σχήμα 5.11: Μέσο ποσοστό επιτυχίας 4^{ης} σειράς αποτελεσμάτων για τα κατηγοριοποιημένα δεδομένα

Όπως έχουμε δει και πιο πάνω, με την αλλαγή στα δεδομένα, έχω καταφέρει να πετύχει την εκπαίδευση του δικτύου, με μέσο ποσοστό επιτυχίας γύρω στο 60%. Παρατήρησα επίσης ότι σε κάθε εκπαίδευση συνήθως ξεχωρίζει μια από τις 2 κατηγορίες, όσο αφορά τα ποσοστά επιτυχίας. Φυσικά η κατηγορία αυτή δεν είναι πάντα σταθερή, αφού σε μια εκπαίδευση μπορεί η κατηγορία των ‘καλοπληρωτών’ να έχει μεγαλύτερα ποσοστά επιτυχίας, ενώ σε μια άλλη, καλύτερα ποσοστά να κατέχει η κατηγορία των ‘κακοπληρωτών’. Σαν τρίτη προσπάθεια μου για επίτευξη μεγαλύτερων ποσοστών επιτυχίας, αποφάσισα να κανονικοποιήσω τα δεδομένα, μετατρέποντας έτσι τις κατηγοριοποιημένες τιμές, σε τιμές μεταξύ του εύρους 0 έως 1. Αυτό θα βοηθούσε στην καλύτερη μάθηση του δικτύου, αφού λόγω της χρήσης της σιγμοειδούς συνάρτησης, όλες οι τιμές εξόδου είναι μεταξύ των τιμών 0 και 1. Επομένως εάν οι τιμές εισόδων κυμαίνονταν στο ίδιο διάστημα, θα βοηθούσαν καλύτερα το δίκτυο να ανταποκριθεί και να εκπαιδευτεί. Όπως παρατηρούμε παρακάτω, η γραφική παράσταση για τα κανονικοποιημένα δεδομένα είναι πιο ομαλή, σε σχέση με αυτήν των κατηγοριοποιημένων δεδομένων. Το σφάλμα εκπαίδευσης καταφέρνει επίσης να μειωθεί ικανοποιητικά. Το γεγονός αυτό μαρτυρεί την καλύτερη εκπαίδευση του δικτύου.

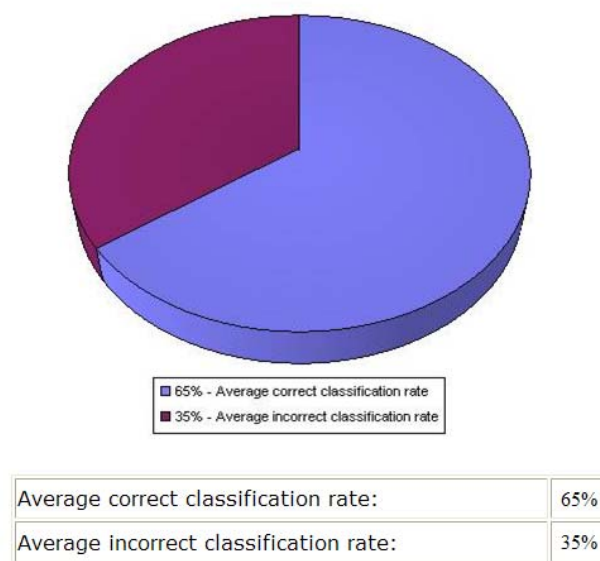


Σχήμα 5.12: Σφάλματα επαλήθευσης και εκπαίδευσης για τα κανονικοποιημένα δεδομένα

Πιο κάτω φαίνεται η 1η σειρά αποτελεσμάτων για τα κανονικοποιημένα δεδομένα. Το δίκτυο έχει καταφέρει να πετύχει 72,5% ποσοστό επιτυχίας για τους 'κακοπληρωτές', σαφώς μεγαλύτερο από αυτό των 'καλοπληρωτών' που φτάνει το 56,79%. Το μέσο ποσοστό επιτυχίας κυμαίνεται στα 65%, το οποίο εμφανίζεται αυξημένο σε σχέση με τα αποτελέσματα των κατηγοριοποιημένων δεδομένων.

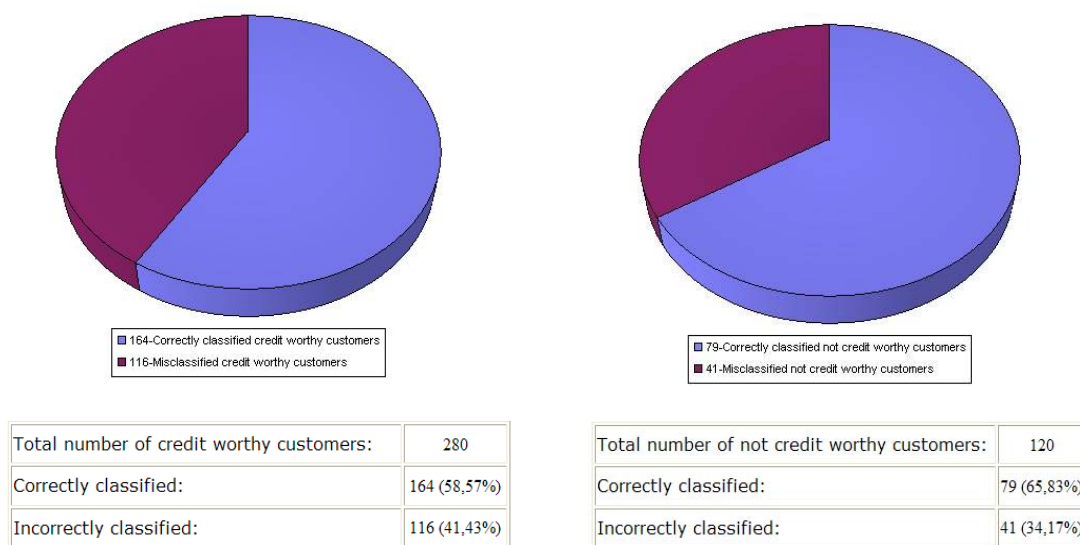


Σχήμα 5.13: 1^η σειρά αποτελεσμάτων για τα κανονικοποιημένα δεδομένα

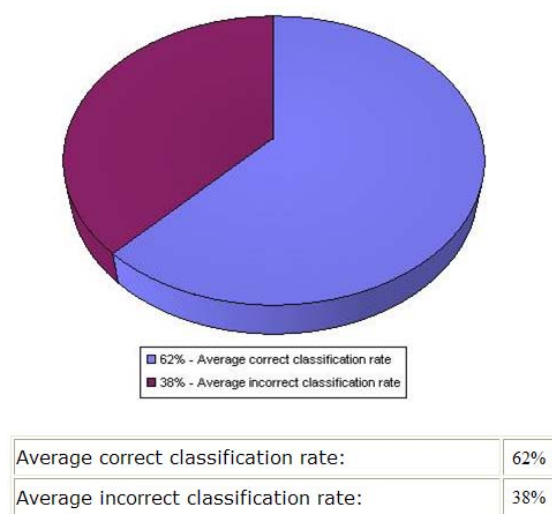


Σχήμα 5.14: Μέσο ποσοστό επιτυχίας 1^{ης} σειράς αποτελεσμάτων για τα κανονικοποιημένα δεδομένα

Στην 2^η σειρά αποτελεσμάτων για τα κανονικοποιημένα δεδομένα όπως φαίνεται παρακάτω, το ποσοστό επιτυχίας για ‘κακοπληρωτές’ ήταν 65,83%, αρκετά μεγαλύτερο από αυτό των ‘καλοπληρωτών’ που ήταν μόνο 58,57%. Έτσι το μέσο ποσοστό επιτυχίας έφτασε μέχρι και το 62%.

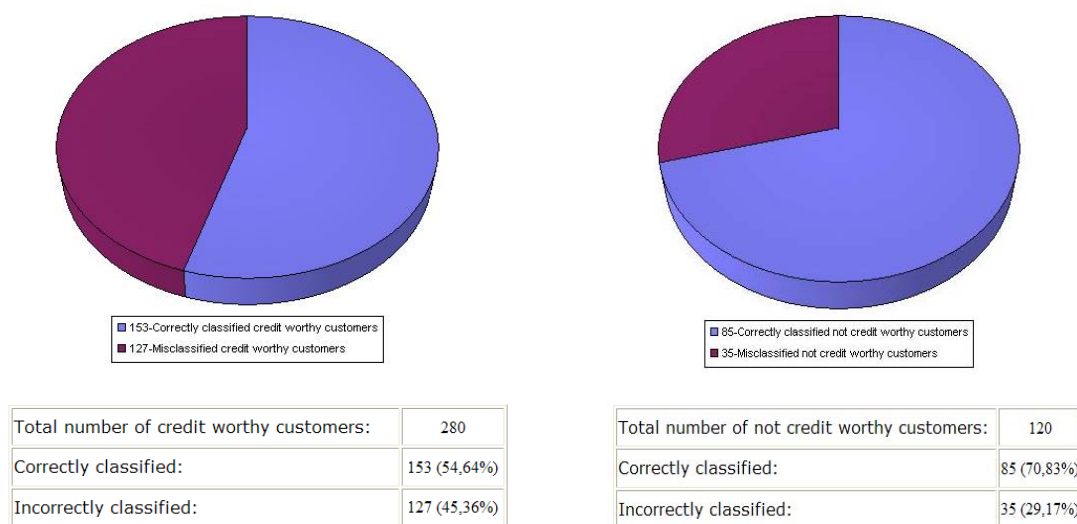


Σχήμα 5.15: 2^η σειρά αποτελεσμάτων για τα κανονικοποιημένα δεδομένα

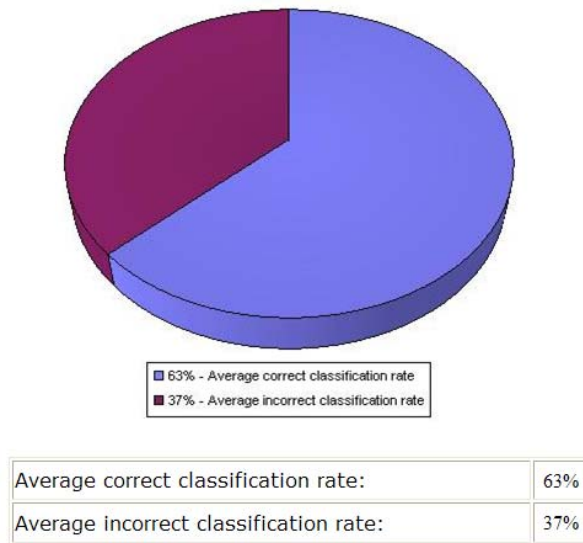


Σχήμα 5.16: Μέσο ποσοστό επιτυχίας 2^{ης} σειράς αποτελεσμάτων για τα κανονικοποιημένα δεδομένα

Στην 3^η σειρά αποτελεσμάτων για τα κανονικοποιημένα δεδομένα όπως φαίνεται παρακάτω, το ποσοστό επιτυχίας για ‘κακοπληρωτές’ ήταν 70,83%, αρκετά μεγαλύτερο από αυτό των ‘καλοπληρωτών’ που ήταν μόνο 54,64%. Έτσι το μέσο ποσοστό επιτυχίας έφτασε μέχρι και το 63%.



Σχήμα 5.17: 3^η σειρά αποτελεσμάτων για τα κανονικοποιημένα δεδομένα



Σχήμα 5.18: Μέσο ποσοστό επιτυχίας 3^{ης} σειράς αποτελεσμάτων για τα κανονικοποιημένα δεδομένα

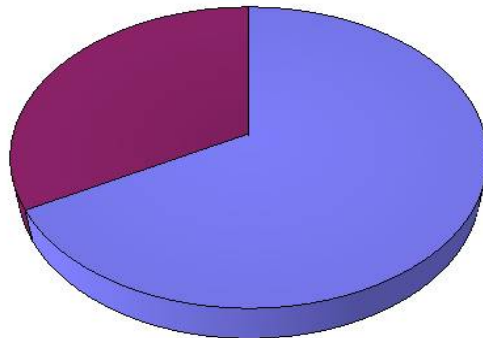
Όπως είδαμε η κανονικοποίηση των δεδομένων, οδήγησε στη βελτιστοποίηση της εκπαίδευσης του δικτύου και της αύξησης του ποσοστού επιτυχίας με ψηλότερο ποσοστό αυτό του 65%. Λόγω αυτού του γεγονότος, στις επόμενες μεθόδους χρησιμοποίησα τα κανονικοποιημένα δεδομένα, αφού οδηγούν σε κάπως καλύτερα αποτελέσματα από τα κατηγοριοποιημένα.

5.2 Αποτελέσματα δεύτερης μεθόδου

Η μέθοδος αυτή έχει ως σκοπό την εξεύρεση ενός βελτιστοποιημένου δικτύου, σε σχέση με αυτά της πρώτης μεθόδου, με όσο μεγαλύτερα ποσοστά επιτυχίας γίνεται. Λόγω του γεγονότος ότι στην προηγούμενη μέθοδο ο χρήστης εισάγει τις παραμέτρους ο ίδιος, δεν μπορεί να είναι σίγουρος για το ποιές τιμές τους είναι οι κατάλληλες. Το πρόβλημα αυτό έρχεται να λύσει η δεύτερη μέθοδος. Στα παρακάτω σχήματα φαίνονται οι 4 σειρές αποτελεσμάτων.

Στην 1η σειρά φαίνεται καθαρά ότι η μέθοδος είναι αποτελεσματική αφού καταφέρνει να πετύχει πολύ ικανοποιητικά αποτελέσματα και συγκεκριμένα της τάξεως του 67%. Χρησιμοποίησε 9 και 10 νευρώνες για το πρώτο και το δεύτερο κρυφό επίπεδο αντίστοιχα, ενώ χρησιμοποίησε κάποια από τα χαρακτηριστικά, αλλά όχι όλα.

• Best Network's classification result:



■ 67% - Average correct classification rate
■ 33% - Average incorrect classification rate

Number of hidden layer 1 neurons: 9

Number of hidden layer 2 neurons: 10

Attributes used:

Balance of current account
Duration in months
Amount of credit
Value of savings
Has been employed by current employer for
Installment in % of available income

Σχήμα 5.19: 1η σειρά αποτελεσμάτων για τη δεύτερη μέθοδο

Στην 2η σειρά η μέθοδος πετυχαίνει και τώρα αποτελέσματα της τάξεως του 67%, αλλά με διαφορετικές τιμές παραμέτρων, γεγονός που αποδεικνύει ότι υπάρχουν αρκετοί συνδυασμοί που μπορούν να πετύχουν καλά αποτελέσματα. Χρησιμοποίησε 20 και 10 νευρώνες για το πρώτο και το δεύτερο κρυφό επίπεδο αντίστοιχα, ενώ χρησιμοποίησε κάποια από τα χαρακτηριστικά, αλλά όχι όλα.

► Best Network's classification result:



■ 67% - Average correct classification rate
■ 33% - Average incorrect classification rate

Number of hidden layer 1 neurons: 20

Number of hidden layer 2 neurons: 10

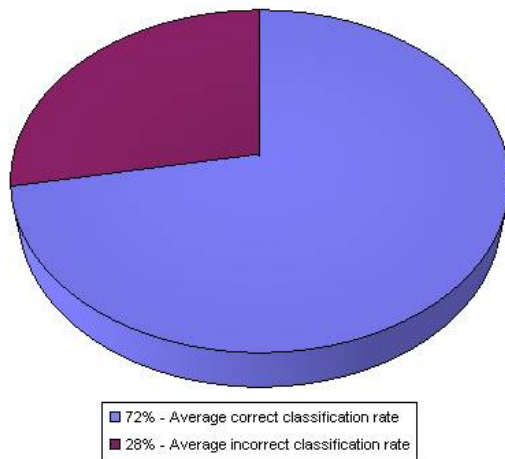
Attributes used:

Has been employed by current employer for
Marital status and sex
Further debtors/guarantors
Living in current household for
Age
Occupation

Σχήμα 5.20: 2η σειρά αποτελεσμάτων για τη δεύτερη μέθοδο

Στην 3η σειρά αποτελεσμάτων το δίκτυο καταφέρνει να πετύχει ποσοστό επιτυχίας της τάξεως του 72% που αποτελεί το ψηλότερο ποσοστό που πέτυχα σε αυτή την εργασία. Χρησιμοποίησε 19 και 7 νευρώνες για το πρώτο και το δεύτερο κρυφό επίπεδο αντίστοιχα, ενώ χρησιμοποίησε κάποια από τα χαρακτηριστικά, αλλά όχι όλα.

Best Network's classification result:



Number of hidden layer 1 neurons: 19

Number of hidden layer 2 neurons: 7

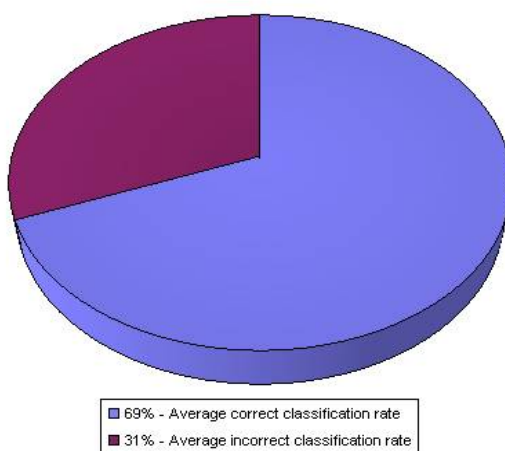
Attributes used:

Balance of current account
Duration in months
Payment of previous credits
Amount of credit
Value of savings
Marital status and sex

Σχήμα 5.21: 3η σειρά αποτελεσμάτων για τη δεύτερη μέθοδο

Στην 4η σειρά αποτελεσμάτων το δίκτυο καταφέρνει να πετύχει ποσοστό επιτυχίας της τάξεως του 69%. Χρησιμοποίησε 20 και 10 νευρώνες για το πρώτο και το δεύτερο κρυφό επίπεδο αντίστοιχα, ενώ χρησιμοποίησε κάποια από τα χαρακτηριστικά, αλλά όχι όλα.

Best Network's classification result:



Number of hidden layer 1 neurons: 20

Number of hidden layer 2 neurons: 10

Attributes used:

Balance of current account
Duration in months
Payment of previous credits
Has been employed by current employer for
Marital status and sex
Further debtors/guarantors

Σχήμα 5.22: 4η σειρά αποτελεσμάτων για τη δεύτερη μέθοδο

Από τα αποτελέσματα αυτής της μεθόδου βλέπουμε ότι έχει αυξηθεί σημαντικά το ποσοστό επιτυχίας. Παρατηρούμε επίσης ότι σημαντικό ρόλο παίζουν οι νευρώνες των δύο κρυφών επιπέδων, αλλά ακόμη σημαντικότερο παίζουν τα δεδομένα, αφού ανάλογα με το ποιά από αυτά χρησιμοποιούνται, τόσο καλύτερα ή χειρότερα αποτελέσματα μπορούμε να έχουμε. Τέλος παρατηρούμε ότι κάποια από τα γνωρίσματα που συμμετέχουν στις εκπαιδεύσεις δικτύων, με υψηλά ποσοστά επιτυχίας, είναι η κατάσταση λογαριασμού, το ποσό του δανείου, η διάρκεια σε μήνες καθώς και η πληρωμή προηγούμενων δανείων.

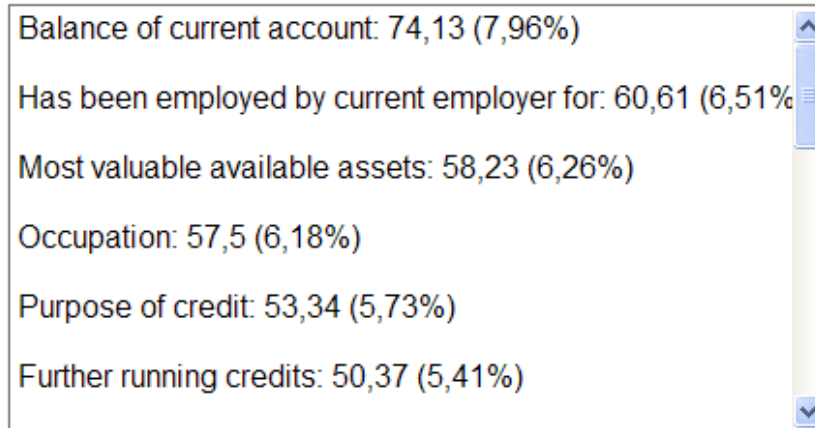
5.3 Αποτελέσματα τρίτης μεθόδου

Σκοπός της μεθόδου αυτής είναι η απομόνωση των 3-4 κυριοτέρων γνωρισμάτων, τα οποία έχουν το μεγαλύτερο βαθμό συμβόλης στην εκπαίδευση το δικτύου. Έχοντας κάνει αρκετές δοκιμές και σε συνδυασμό με τα αποτελέσματα της μεθόδου του γενετικού αλγορίθμου, κατέληξα στο συμπέρασμα ότι οι παράγοντες που παίζουν ρόλο στην εκπαίδευση του δικτύου, μπορούν να ποικίλουν, όμως κάποιοι από αυτούς εμφανίζονται συχνότερα από άλλους γι'αυτό και τους ξεχώρισα ως τους σημαντικότερους παράγοντες. Αυτοί είναι:

- Σκοπός δανείου
- Ποσό δανείου
- Διάρκεια
- Πληρωμή προηγούμενων δανείων
- Κατάσταση λογαριασμού
- Περιουσιακά στοιχεία

Στην επόμενη σελίδα φαίνονται στιγμιότυπα από τις λίστες των γνωρισμάτων.

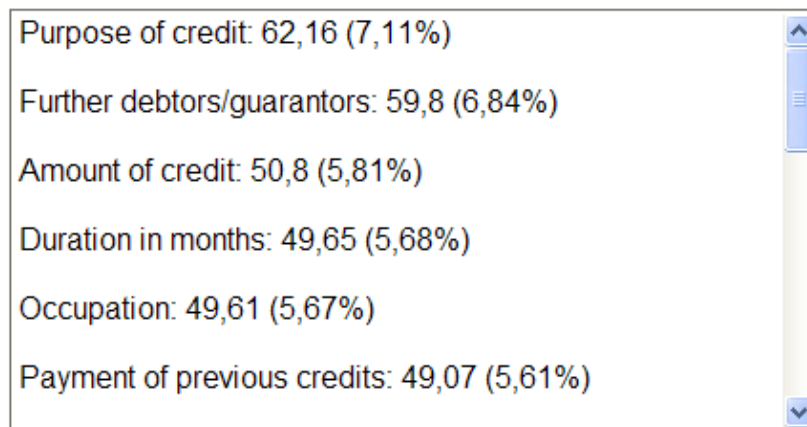
Inputs sensitivity analysis



Balance of current account:	74,13 (7,96%)
Has been employed by current employer for:	60,61 (6,51%)
Most valuable available assets:	58,23 (6,26%)
Occupation:	57,5 (6,18%)
Purpose of credit:	53,34 (5,73%)
Further running credits:	50,37 (5,41%)

Σχήμα 5.23: Λίστα γνωρισμάτων με παράγοντα συμβολής τους

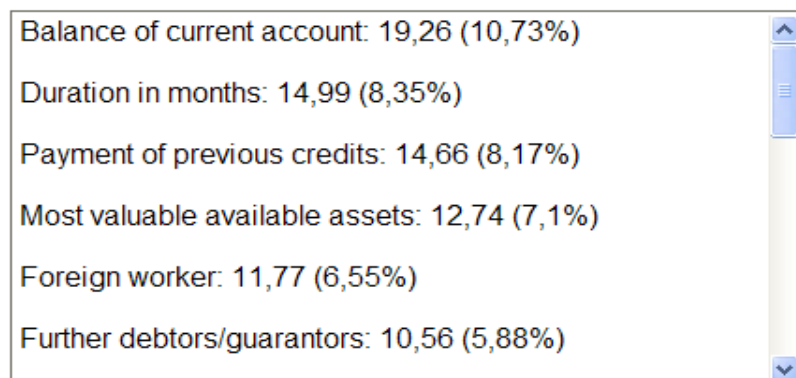
Inputs sensitivity analysis



Purpose of credit:	62,16 (7,11%)
Further debtors/guarantors:	59,8 (6,84%)
Amount of credit:	50,8 (5,81%)
Duration in months:	49,65 (5,68%)
Occupation:	49,61 (5,67%)
Payment of previous credits:	49,07 (5,61%)

Σχήμα 5.24: Λίστα γνωρισμάτων με παράγοντα συμβολής τους

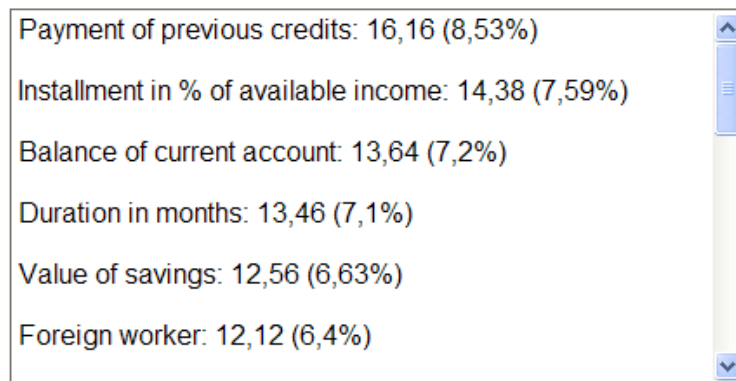
Inputs sensitivity analysis



Balance of current account:	19,26 (10,73%)
Duration in months:	14,99 (8,35%)
Payment of previous credits:	14,66 (8,17%)
Most valuable available assets:	12,74 (7,1%)
Foreign worker:	11,77 (6,55%)
Further debtors/guarantors:	10,56 (5,88%)

Σχήμα 5.25: Λίστα γνωρισμάτων με παράγοντα συμβολής τους

Inputs sensitivity analysis



Payment of previous credits: 16,16 (8,53%)
Installment in % of available income: 14,38 (7,59%)
Balance of current account: 13,64 (7,2%)
Duration in months: 13,46 (7,1%)
Value of savings: 12,56 (6,63%)
Foreign worker: 12,12 (6,4%)

Σχήμα 5.26: Λίστα γνωρισμάτων με παράγοντα συμβολής τους

Κεφάλαιο 6

Συμπεράσματα

6.1 Γενικά συμπεράσματα	60
6.2 Μελλοντική εργασία	60

6.1 Γενικά συμπεράσματα

Σε γενικές γραμμές κατάφερα να πετύχω αρκετά καλά αποτελέσματα, με ψηλότερο το ποσοστό επιτυχίας της τάξεως του 72%. Σε σύγκριση με το άρθρο 'Credit scoring using the hybrid neural discriminant technique' των Tian-Shyug Lee, Chih-Chou Chiu, Chi-Jie Lu, και I-Fei Chend πιστεύω τα αποτελέσματα μου ήταν πολύ καλά, αφού οι συγγραφείς στην περίπτωση τους κατάφεραν με απλό νευρωνικό δίκτυο να πετύχουν αποτελέσματα μέχρι και 73,7%, όχι πολύ πιο πάνω από τα δικά μου. Κατάφερα να φτάσω στο συμπέρασμα ότι τα κανονικοποιημένα δεδομένα μπορούν αν είναι πιο αποτελεσματικά, όσο αφορά την εκπαίδευση του δικτύου. Ακόμη κατάφερα να δείξω ότι ο συνδυασμός συγκεκριμένων παραμέτρων του δικτύου και κάποιων γνωρισμάτων μπορεί να οδηγήσει σε καλύτερα αποτελέσματα. Η δεύτερη μέθοδος συγκεκριμένα μπορεί να δώσει αρκετούς συνδυασμούς για την επίτευξη όσο το δυνατό βέλτιστων αποτελεσμάτων. Τέλος κατάφερα να απομονώσω κάποια από τα γνωρίσματα, τα οποία έχουν συνήθως το μεγαλύτερο ποσοστό συμβολής στην εκπαίδευση του δικτύου και παίζουν σημαντικότερο ρόλο από τα άλλα, γεγονός αρκετά σημαντικό.

6.2 Μελλοντική εργασία

Όσο αφορά τη μελλοντική εργασία, η δεύτερη μέθοδος του συστήματος θα μπορούσε να δέχεται ως πιθανές λύσεις, χρωμοσώματα που περιέχουν επιπλέον και τις μεταβλητές της ορμής και του ρυθμού μάθησης. Με αυτόν τον τρόπο ο γενετικός αλγόριθμος θα προσπαθούσε να βελτιστοποιήσει όλες τις μεταβλητές του δικτύου,

χωρίς να αφήνει την επιλογή των τιμών, ορισμένων από αυτών, στο χρήστη. Επίσης θα μπορούσε να ληφθεί υπόψη στην αξιολόγηση των λύσεων, εκτός από τα αποτελέσματα πρόβλεψης, και ο βαθμός πολυπλοκότητας του δικτύου, έτσι ώστε να επιλέγεται το αποτελεσματικότερο αλλά και λιγότερο πολύπλοκο δίκτυο.

Βιβλιογραφία

- [1] Tian-Shyug Lee, Chih-Chou Chiu, Chi-Jie Lu and I-Fei Chen, “Credit scoring using the hybrid neural discriminant technique”, Expert Systems with Applications 23, 245–254, 2002
- [2] M. B. Yobas, J. N. Crook and P. Ross, “Credit scoring using neural and evolutionary techniques”, IMA Journal of Mathematics Applied in Business and Industry (2000) 11, 111-125
- [3] Li Zhan, Xu ji-Sheng and Xu Min, “ANN-GA Approach of Credit Scoring for Mobile Users”, Proceedings of the 2004 IEEE, Conference on Cybernetics and Intelligent Systems, Singapore, 1-3 December, 2004
- [4] Helen McNab and Anthea Wynn, “Principles and practice of consumer credit risk management”, 25 October, 2004
- [5] Lyn C. Thomas, David B. Edelman and Jonathan N. Crook, “Credits scoring and its applications”, 31 January 2002
- [6] Datasets at the Department of Statistics, University of Munich

Παράρτημα Α

Πιο κάτω υπάρχει ο πίνακας στον οποίο φαίνονται οι κατηγορίες κάθε γνωρίσματος και οι τιμές που παίρνουν, σύμφωνα με την ιστοσελίδα του πανεπιστημίου του Μόναχου και οι οποίες χρησιμοποιήθηκαν στην εργασία αυτή.

Πίνακας Α-1: Περιγραφή δεδομένων

Description	Categories	Value
Balance of current account	No balance or debit	2
	Less than 200 Euros	3
	More than 200 Euros	4
	No running account	1
Duration in months	≤ 6	10
	$6 < \dots \leq 12$	9
	$12 < \dots \leq 18$	8
	$18 < \dots \leq 24$	7
	$24 < \dots \leq 30$	6
	$30 < \dots \leq 36$	5
	$36 < \dots \leq 42$	4
	$42 < \dots \leq 48$	3
	$48 < \dots \leq 54$	2
	> 54	1
Payment of previous credits	No previous credits	2
	Paid back previous credits at this bank	4
	No problems with current credits at this bank	3

	Hesitant payment of previous credits	0
	There are further credits running but at other banks	1
Purpose of credit	new car	1
	used car	2
	items of furniture	3
	radio / television	4
	household appliances	5
	repair	6
	education	7
	vacation	8
	retraining	9
	business	10
	other	0
Amount of credit	≤ 500	10
	$500 < \dots \leq 1000$	9
	$1000 < \dots \leq 1500$	8
	$1500 < \dots \leq 2500$	7
	$2500 < \dots \leq 5000$	6
	$5000 < \dots \leq 7500$	5
	$7500 < \dots \leq 10000$	4
	$10000 < \dots \leq 15000$	3
	$15000 < \dots \leq 20000$	2
	> 20000	1
Value of savings	Less than 1000 Euros	2
	$1000 \leq \dots < 5000$ Euros	3

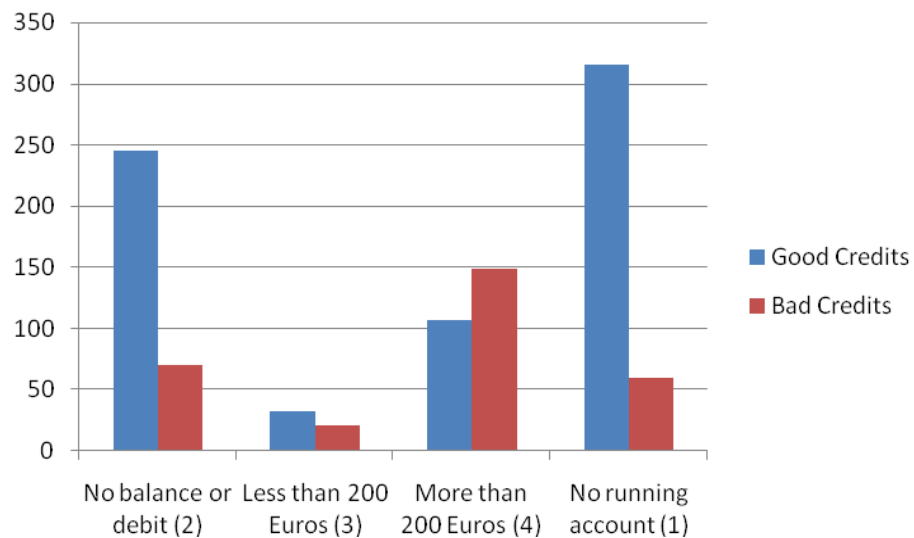
	5000 <= ... < 10000 Euros	4
	More than 10000 Euros	5
	No savings	1
Has been employed by current employer for	Unemployed	1
	<= 1 year	2
	1 <= ... < 4 years	3
	4 <= ... < 7 years	4
	>= 7 years	5
Installment in % of available income	>= 35	1
	25 <= ... < 35	2
	20 <= ... < 25	3
	< 20	4
Marital Status / Sex	male: divorced / living apart	1
	female: divorced / living apart / married & Male: single	2
	male: single	2
	male: married / widowed	3
	female: single	4
Further debtors / Guarantors	none	1
	Co-Applicant	2
	Guarantor	3
Living in current household for	< 1 year	1
	1 <= ... < 4 years	2
	4 <= ... < 7 years	3
	>= 7 years	4

Most valuable available assets	Ownership of house or land	4
	Savings contract with a building society / Life insurance	3
	Car / Other	2
	Not available / no assets	1
Age in years	0 <= ... <= 25	1
	26 <= ... <= 39	2
	40 <= ... <= 59	3
	60 <= ... <= 64	5
	>= 65	4
Further running credits	At other banks	1
	At department store or mail order house	2
	No further running credits	3
Type of apartment	Rented flat	2
	Owner-occupied flat	3
	Free apartment	1
Number of previous credits at this bank	One	1
	Two or three	2
	Four or five	3
	Six or more	4
Occupation	Unemployed / unskilled with no permanent residence	1
	Unskilled with permanent residence	2
	Skilled worker / skilled employee / minor civil servant	3
	Executive / self-employed / higher civil servant	4

Number of persons entitled to maintenance	0 to 2	2
	3 and more	1
Telephone	No	1
	Yes	2
Foreign worker	Yes	1
	No	2

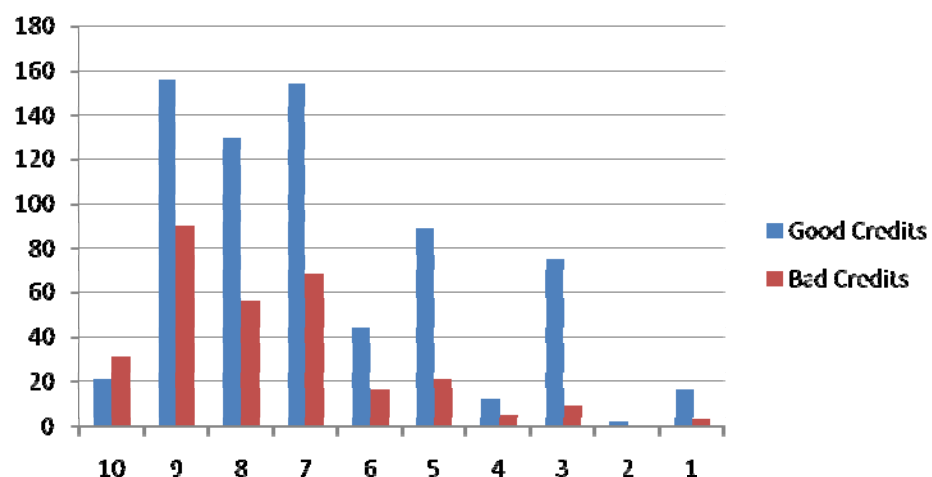
Πιο κάτω φαίνονται οι κατανομές των αξιόπιστων και μη δειγμάτων του συνόλου δεδομένων, σε σχέση με κάθε ένα γνώρισμα ξεχωριστά. Να σημειώσω ότι ο κάθετος άξονας αναπαριστά τον αριθμό δειγμάτων. Σε κάποιες γραφικές παραστάσεις στον οριζόντιο άξονα φαίνονται μόνο οι τιμές των κατηγοριών του, για λόγους καλαισθησίας. Το όνομα των κατηγοριών φαίνεται στον πίνακα Α-1 του ίδιου παραρτήματος.

1. Κατάσταση λογαριασμού



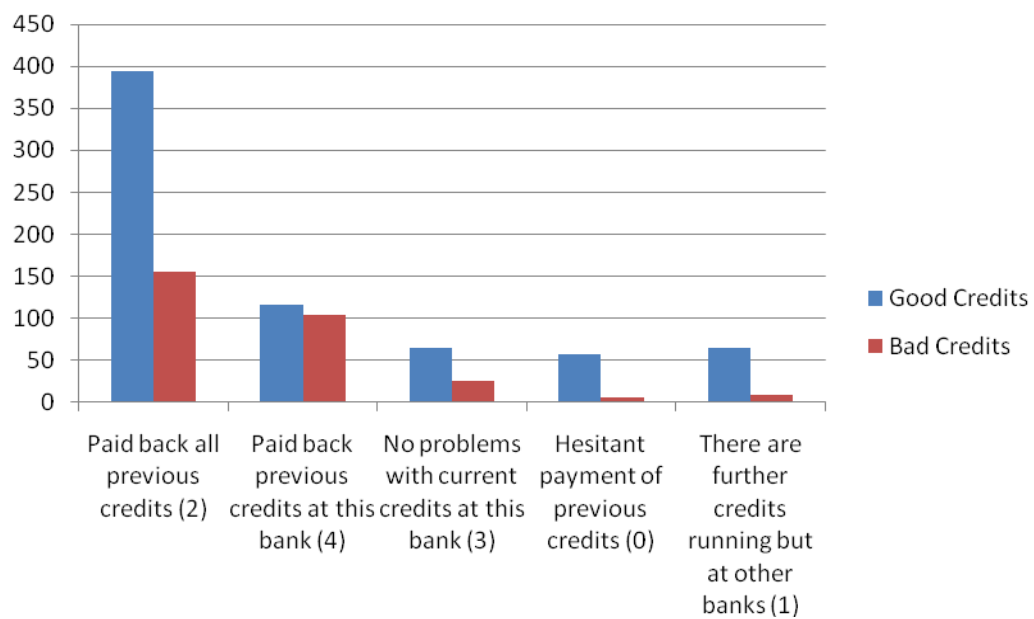
Γραφική παράσταση Α-1: Κατανομή δειγμάτων σε σχέση με την κατάσταση λογαριασμού

2. Διάρκεια



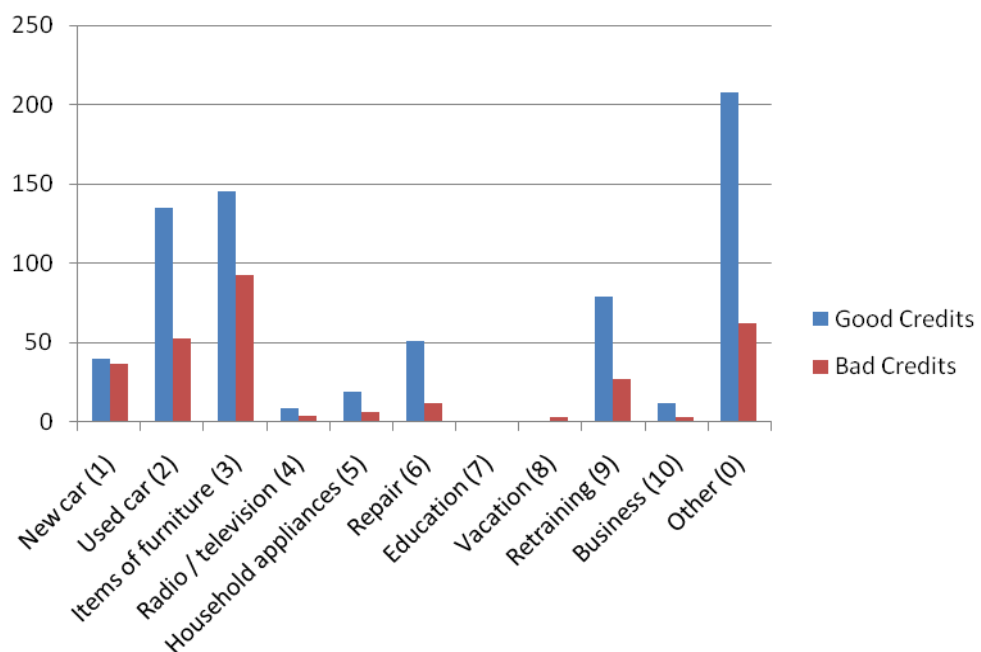
Γραφική παράσταση Α-2: Κατανομή δειγμάτων σε σχέση με διάρκεια

3. Πληρωμή προηγούμενων δανείων



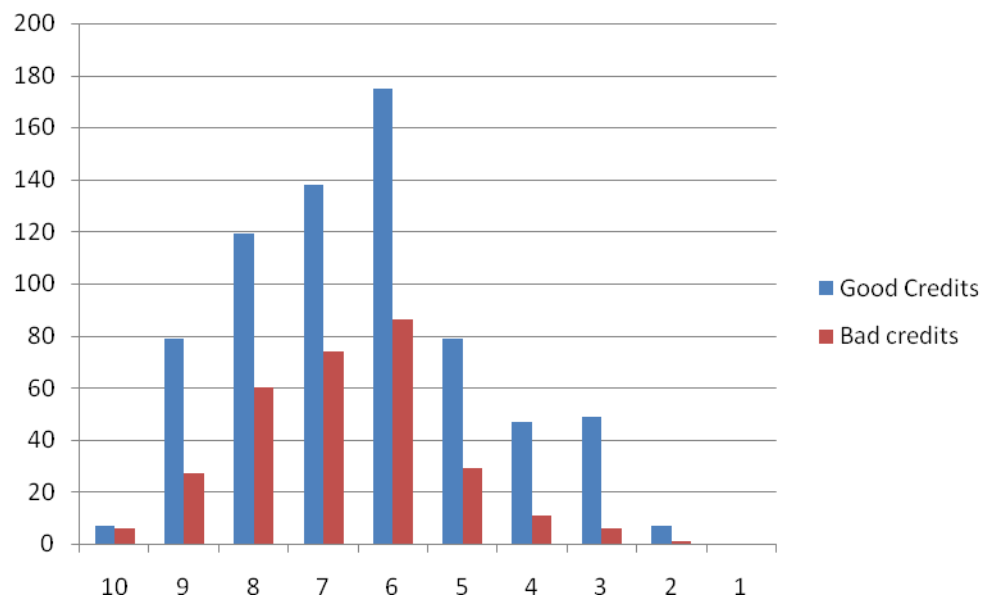
Γραφική παράσταση Α-3: Κατανομή δειγμάτων σε σχέση με την πληρωμή προηγούμενων δανείων

4. Σκοπός δανείου



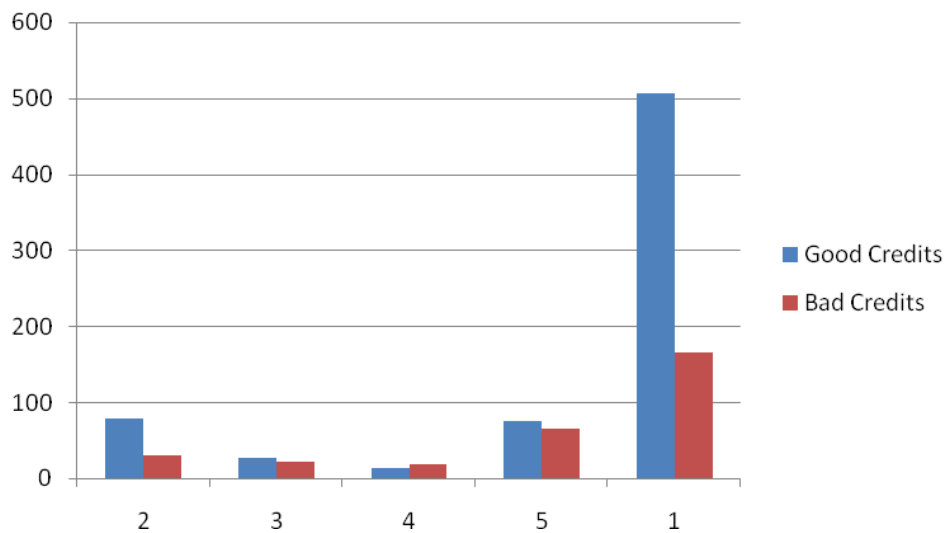
Γραφική παράσταση Α-4: Κατανομή δειγμάτων σε σχέση με το σκοπό δανείου

5. Ποσό δανείου



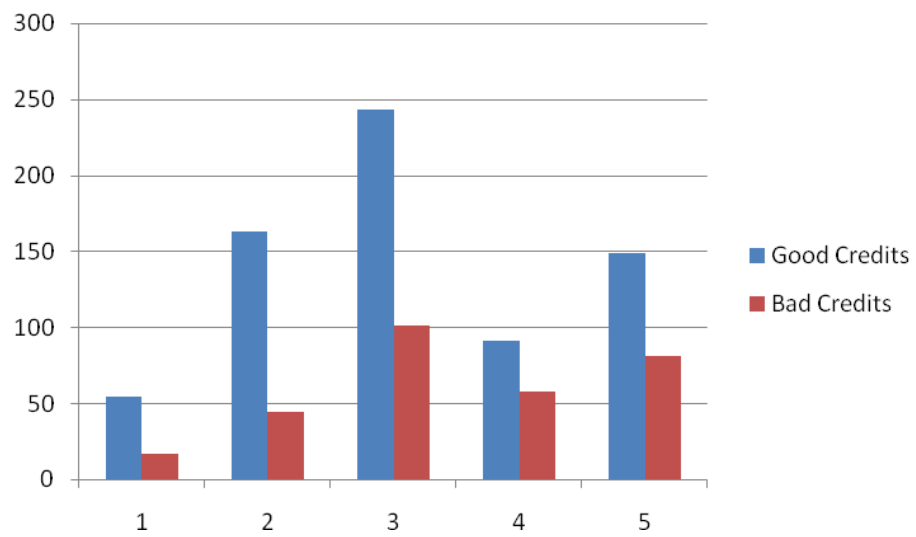
Γραφική παράσταση Α-5: Κατανομή δειγμάτων σε σχέση με το ποσό δανείου

6. Αξία μετοχών και οικονομιών



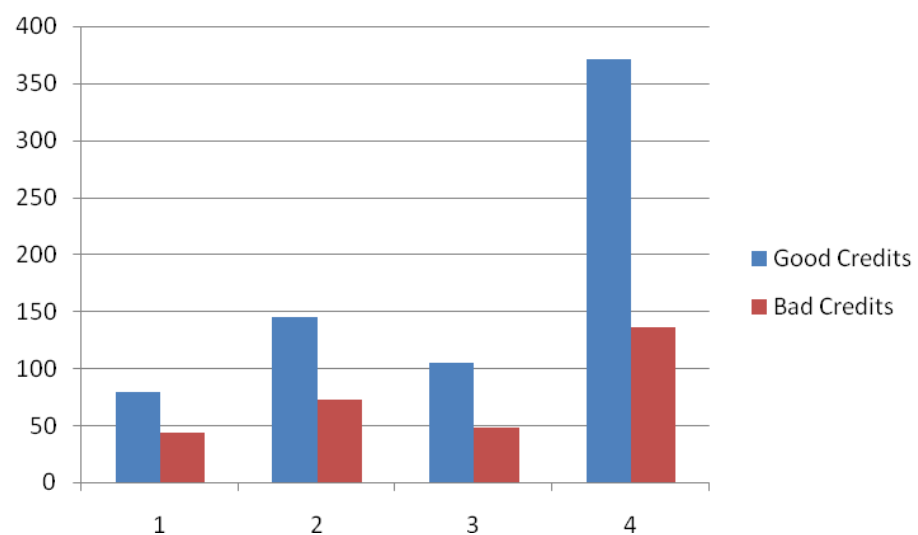
Γραφική παράσταση Α-6: Κατανομή δειγμάτων σε σχέση με την αξία μετοχών και οικονομιών

7. Χρόνος που εργάζεται στον τωρινό εργοδότη



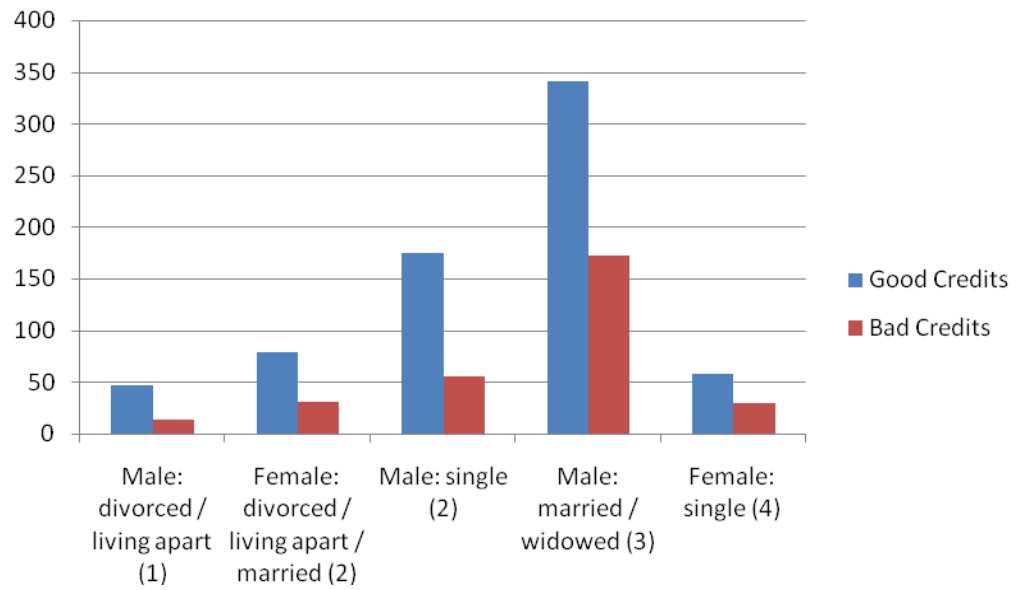
Γραφική παράσταση Α-7: Κατανομή δειγμάτων σε σχέση με χρόνο εργασίας στον τωρινό εργοδότη

8. Δόση, υπο μορφή ποσοστού του διαθέσιμου εισοδήματος



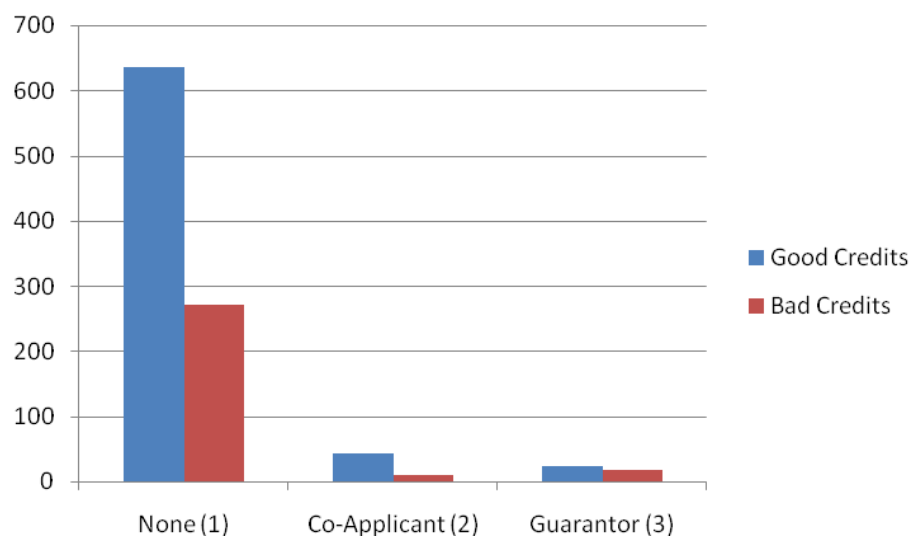
Γραφική παράσταση Α-8: Κατανομή δειγμάτων σε σχέση με τη δόση, ως ποσοστό του διαθέσιμου εισοδήματος

9. Οικογενειακή κατάσταση / Φύλο



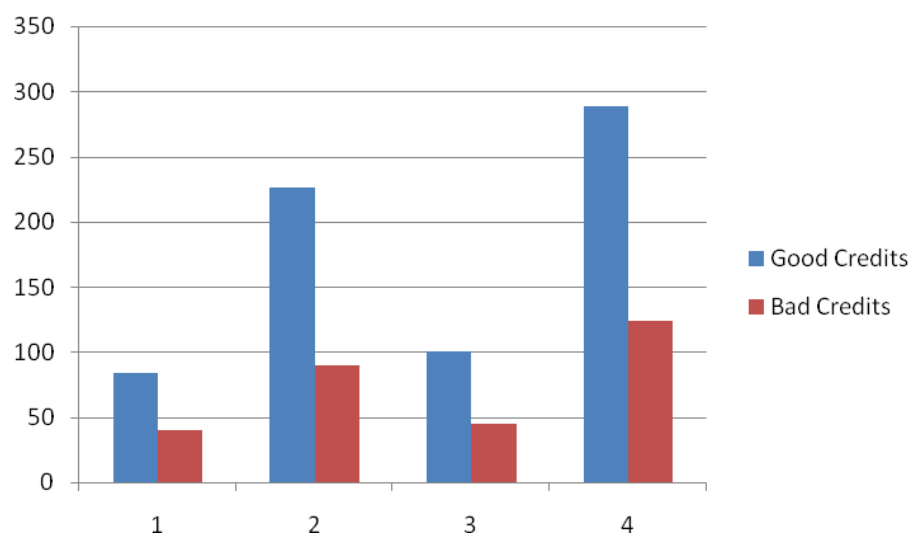
Γραφική παράσταση Α-9: Κατανομή δειγμάτων σε σχέση με την οικογενειακή κατάσταση και το φύλο

10. Εγγυητές



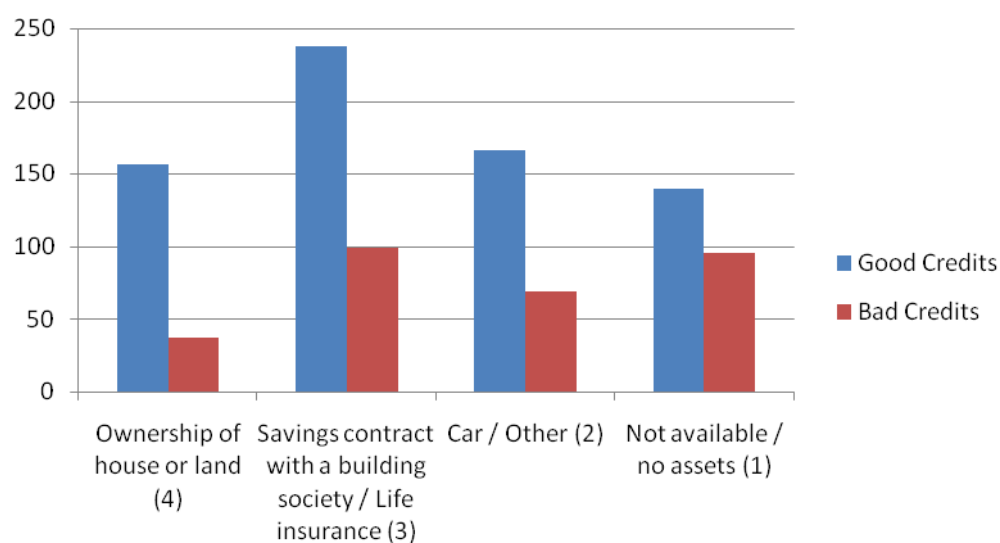
Γραφική παράσταση Α-10: Κατανομή δειγμάτων σε σχέση με τους εγγυητές

11. Χρόνια ζωής στο τωρινό νοικοκυριό



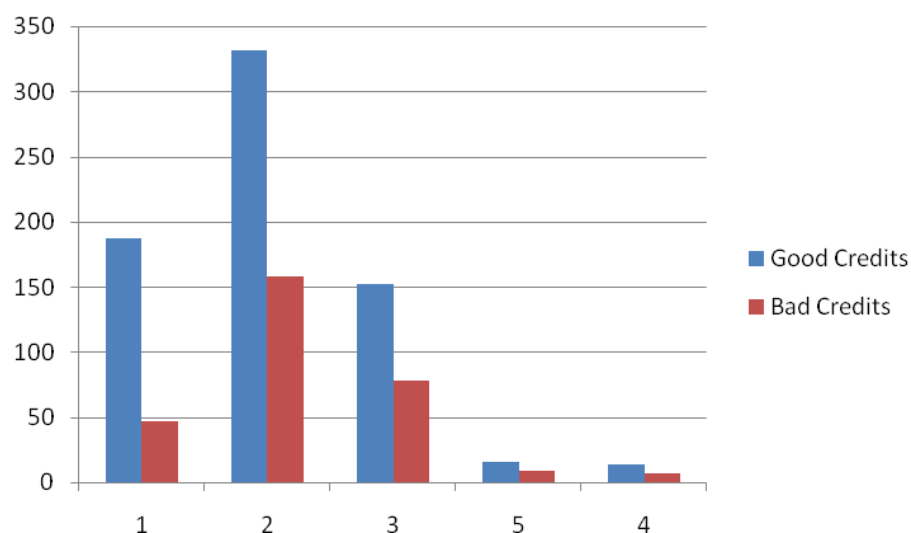
Γραφική παράσταση A-11: Κατανομή δειγμάτων σε σχέση με τα χρόνια ζωής στο τωρινό νοικοκυριό/σπίτι

12. Περιουσιακά στοιχεία



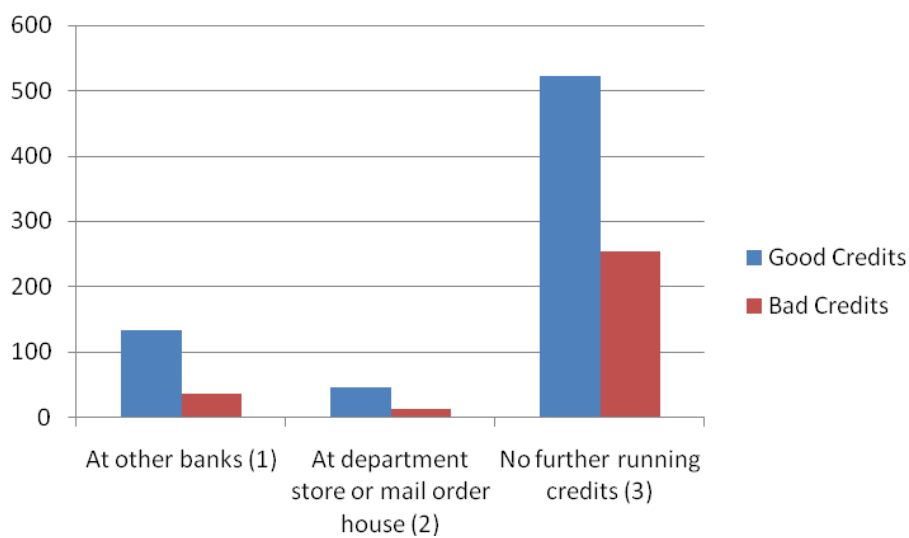
Γραφική παράσταση A-12: Κατανομή δειγμάτων σε σχέση με τα περιουσιακά στοιχεία

13. Ηλικία



Γραφική παράσταση A-13: Κατανομή δειγμάτων σε σχέση με την ηλικία

14. Επιπλέον ενεργά δάνεια



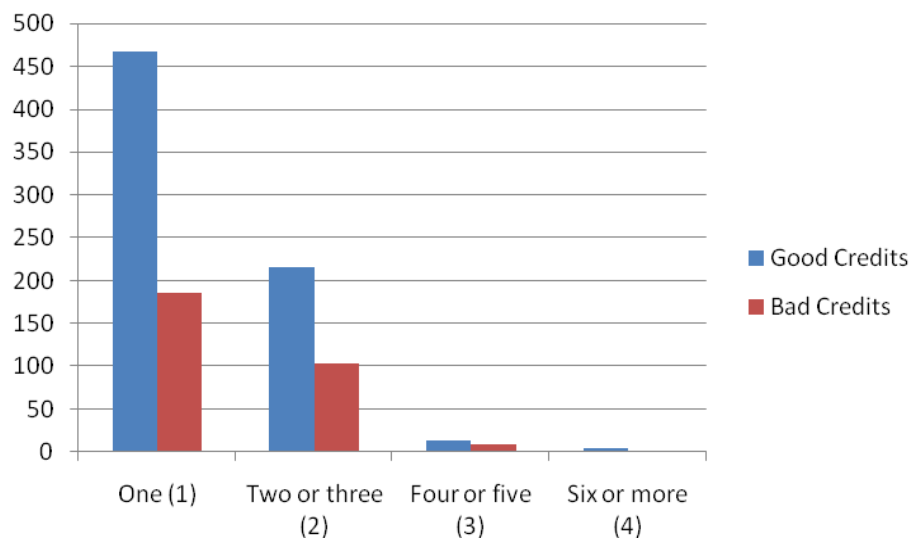
Γραφική παράσταση A-14: Κατανομή δειγμάτων σε σχέση με τα επιπλέον ενεργά δάνεια

15. Τύπος σπιτιού



Γραφική παράσταση A-15: Κατανομή δειγμάτων σε σχέση με τον τύπο σπιτιού

16. Αριθμός προηγούμενων δανείων σε αυτή την τράπεζα



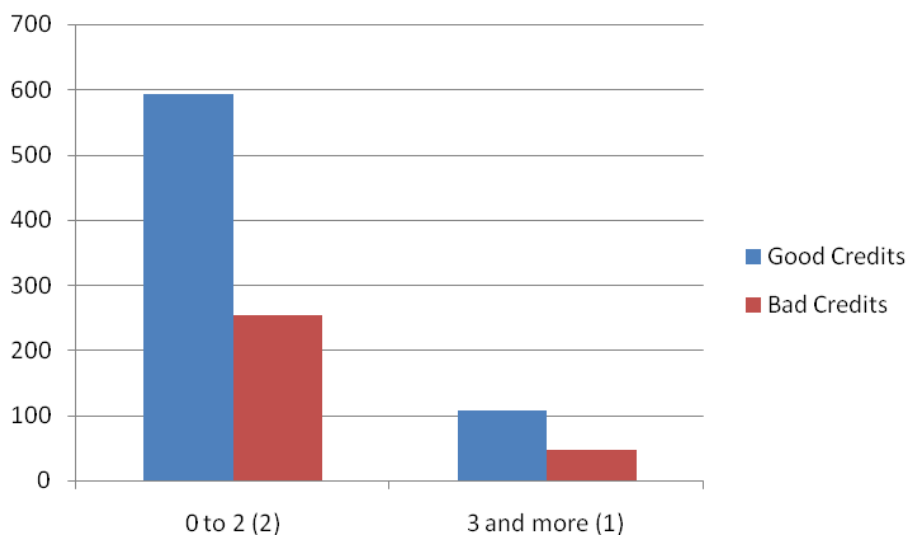
Γραφική παράσταση A-16: Κατανομή δειγμάτων σε σχέση με τον αριθμό των προηγούμενων δανείων στη συγκεκριμένη τράπεζα

17. Επάγγελμα



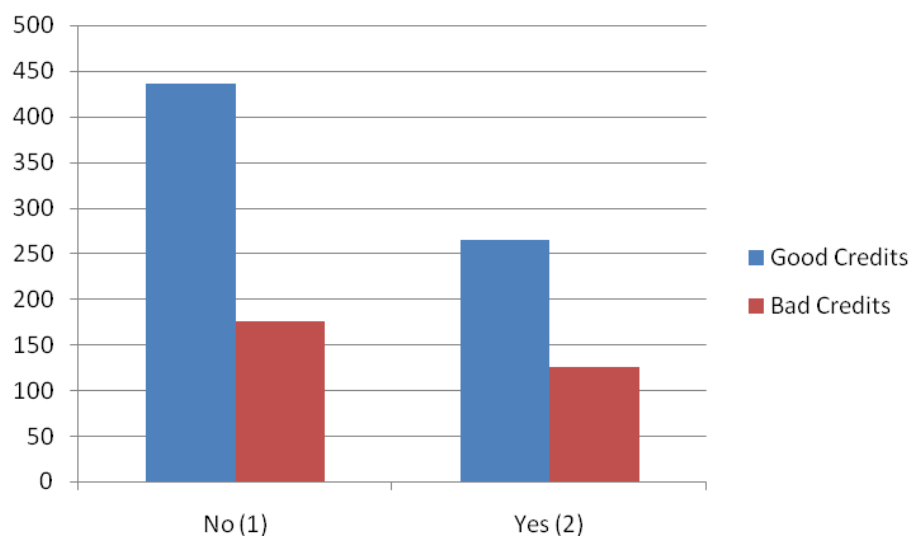
Γραφική παράσταση A-17: Κατανομή δειγμάτων σε σχέση με το επάγγελμα

18. Αριθμός ατόμων προς συντήρηση



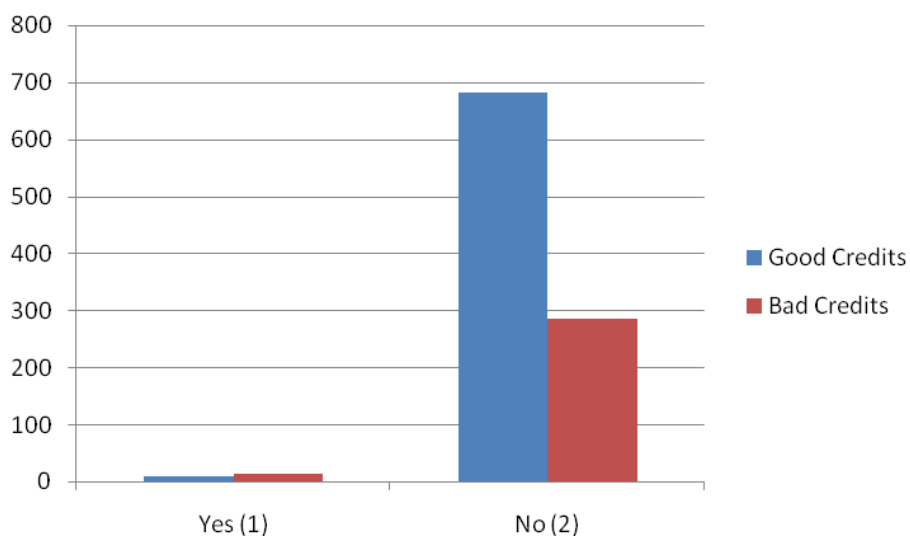
Γραφική παράσταση A-18: Κατανομή δειγμάτων σε σχέση με τον αριθμό ατόμων προς συντήρηση

19. Τηλέφωνο



Γραφική παράσταση A-19: Κατανομή δειγμάτων σε σχέση με την κατοχή συσκευής τηλεφώνου

20. Ξένος εργάτης



Γραφική παράσταση A-20: Κατανομή δειγμάτων σε σχέση με την ιδιότητα ξένος εργάτης