ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Νευρωνικά Δίκτυα Μακράς και Βραχείας Μνήμης για Πρόβλεψη Τραπεζικών Διαδικασιών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΕΛΙΝΑ ΜΑΗ

Επιβλέπων : Μέντζας Γρηγόρης
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2021

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Νευρωνικά Δίκτυα Μακράς και Βραχείας Μνήμης για Πρόβλεψη Τραπεζικών Διαδικασιών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΕΛΙΝΑ ΜΑΗ

Επιβλέπων : Γρηγόρης Μέντζας
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 19η Ιουλίου 2021.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Γρηγόρης Μέντζας	Δημήτριος Ασκούνης	Χάρης Δούκας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Αν. Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2021

(Υπογραφή)

.....

ΜΕΛΙΝΑ ΜΑΗ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η πρόβλεψη της συμπεριφοράς της επιχειρηματικής διαδικασίας είναι μια σημαντική πτυχή της διαχείρισης επιχειρησιακών διαδικασιών (Business Process Management) και οι τεχνικές βαθιάς μάθησης (deep learning) έχουν βρει πρόσφατα εφαρμογές στον τομέα. Οι προβλεπτικές μέθοδοι παρακολούθησης επιχειρηματικών διαδικασιών εκμεταλλεύονται αρχεία καταγραφής ολοκληρωμένων γεγονότων (event logs) για να κάνουν προβλέψεις σχετικά με την εκτέλεση των περιπτώσεων (cases) αυτών. Οι υφιστάμενες μέθοδοι στο χώρο αυτό είναι προσαρμοσμένες για συγκεκριμένες εργασίες και σύνολα δεδομένων. Αυτή η διπλωματική εργασία διερευνά τον τρόπο χρήσης τεχνικών βαθιάς μάθησης και συγκεκριμένα επαναλαμβανόμενων νευρωνικών δικτύων με αρχιτεκτονική μακράς και βραχείας μνήμης (Long-Short-Term Memory (LSTM)) για να προβλέψει την επόμενη δραστηριότητα σε μια επιχειρηματική διαδικασία. Η προσέγγιση αξιολογείται σε ένα πραγματικό σύνολο δεδομένων του τραπεζικού τομέα και τα αποτελέσματα δείχνουν ότι η πρόβλεψη της επόμενης δραστηριότητας είναι αποδεκτή σύμφωνα με τη βιβλιογραφία του πεδίου.

Λέξεις Κλειδιά: Εξόρυξη διαδικασιών, Πρόβλεψη διαδικασιών, Βαθιά μηχανική μάθηση, Νευρωνικά δίκτυα μακράς και βραχείας μνήμης, Αρχεία καταγραφής γεγονότων, Επιχειρησιακές διαδικασίες

Abstract

Predicting business process behaviour is an important aspect of business process management, and deep learning techniques have recently found applications in the field. Predictive business process monitoring methods exploit logs of completed cases to make predictions about running cases thereof. Existing methods in the space our customized for specific tasks and datasets. This thesis investigates how to use deep learning techniques and specifically recurrent neural networks with Long-Short-Term Memory (LSTM) architecture to predict the next event in a business process. The approach is evaluated on a real dataset of the banking sector and results indicate that the prediction of the next activity is acceptable according to the literature of the domain.

Keywords: Process mining, Process prediction, Deep learning, LSTM, Event Log, Business Process

Ευχαριστίες

.

Με την ολοκλήρωση της διπλωματικής μου εργασίας και των προπτυχιακών μου σπουδών θα ήθελα να ευχαριστήσω τους ανθρώπους που με βοήθησαν και με στήριξαν στην προσπάθεια μου αυτή.

Καταρχάς, θα ήθελα να ευχαριστήσω θερμά τον καθηγητή κ.Γρηγόρη Μέντζα για την εμπιστοσύνη που μου έδειξε με την ανάθεση αυτής της διπλωματικής εργασίας. Επίσης, θέλω να ευχαριστήσω ιδιαίτερα τους διδακτορικούς ερευνητές Κατερίνα Λεπενιώτη και Αλέξανδρο Βουσδέκη για την καθοδήγηση και την ενθάρρυνση που μου παρείχαν καθ' όλη τη διάρκεια της εκπόνησης της εργασίας αυτής. Οι συμβουλές, η αφοσίωση και ο οργανωμένος τρόπος λειτουργίας τόσο του καθηγητή αλλά και των διδακτορικών ερευνητών έκαναν αυτό το τελευταίο κομμάτι των σπουδών μου ανεπανάληπτο και νιώθω ευγνώμων για την ευκαιρία που είχα να συνεργαστώ μαζί τους.

Επιπλέον, θα ήθελα να ευχαριστήσω όλους τους φίλους μου για την υποστήριξη και όλους τους συμφοιτητές με τους οποίους συμπορεύτηκα στο Πολυτεχνείο και σημάδεψαν την φοιτητική μου πορεία με όμορφες στιγμές και αναμνήσεις.

Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου, Τασούλα και Μάριο, καθώς και όλη μου την οικογένεια, για την ανιδιοτελή στήριξη, την αγάπη και τα εφόδια που μου έδωσαν όλα αυτά τα χρόνια.

Μελίνα Μάη

Αθήνα, 19η Ιουλίου 2021

Πίνακας περιεχομένων

Πίνακας Πινάκων

Πίνακας Εικόνων

1

Εισαγωγή

1.1 Προβλέψεις επόμενων δραστηριοτήτων στις επιχειρηματικές διαδικασίες

Σε μια οικονομία βασισμένη στη γνώση, δημόσιοι και ιδιωτικοί οργανισμοί απαιτούν σωστή διαχείριση των στοιχείων γνώσης τους για να διατηρήσουν ένα ανταγωνιστικό πλεονέκτημα σε παγκόσμιες αγορές ή σε κυβερνητικές υπηρεσίες. Σε αυτό το πλαίσιο, το Business Process Management (BPM) θεωρείται βασικό συστατικό για τη διαχείριση του κύκλου ζωής των επιχειρηματικών διαδικασιών που ενορχηστρώνουν τις δραστηριότητες που εκτελούνται σε οργανισμούς, καθώς και τους πόρους (ανθρώπους, ρομπότ ή συστήματα πληροφοριών) που εκτελούν τέτοιες δραστηριότητες . Μια επιχειρηματική διαδικασία αποτελείται από ένα σύνολο δραστηριοτήτων που εκτελούνται με συντονισμένο τρόπο σε έναν οργανισμό και έχουν τουλάχιστον έναν συσχετισμένο επιχειρηματικό στόχο. Η τυπική γλώσσα για τη μοντελοποίηση επιχειρηματικών διαδικασιών είναι το BPMN (Business Process Modeling Notation) . Τα συστήματα πληροφοριών διαδραματίζουν σημαντικό ρόλο στη διαχείριση των επιχειρηματικών διαδικασιών, επειδή ένας μεγάλος αριθμός των δραστηριοτήτων που εκτελούν οι οργανισμοί υποστηρίζονται από αυτά. Οι πληροφορίες καταχωρούνται σε αρχεία καταγραφής γεγονότων, καθιστώντας τις διαθέσιμες σε σχεδόν πραγματικό χρόνο με βάση ασύγχρονα συμβάντα και εφαρμογές σε επιχειρηματικό επίπεδο που είναι σε θέση να

χρησιμοποιούν πληροφορίες υψηλού επιπέδου για διάφορους σκοπούς, όπως διάγνωση, δείκτες απόδοσης ή ιχνηλασιμότητα. Σε αυτό το πλαίσιο, η πρόβλεψη της συμπεριφοράς μιας επιχειρηματικής διαδικασίας, δηλαδή η εκμετάλλευση των αρχείων καταγραφής γεγονότων για την πραγματοποίηση προβλέψεων σχετικά με την εκτέλεση των δραστηριοτήτων, αποτελεί βασική πτυχή προκειμένου να παρέχει πολύτιμη συμβολή στον προγραμματισμό και την κατανομή πόρων . Υπάρχει λοιπόν, η ανάγκη βελτίωσης και υποστήριξης επιχειρηματικών διαδικασιών σε ανταγωνιστικά και ταχέως μεταβαλλόμενα περιβάλλοντα.

Οι τεχνικές εξόρυξης διεργασιών είναι ικανές να εξάγουν γνώσεις από αρχεία καταγραφής συμβάντων, συνήθως διαθέσιμα σε συστήματα πληροφοριών. Αυτές οι τεχνικές παρέχουν νέα μέσα για την ανακάλυψη, την παρακολούθηση και τη βελτίωση επιχειρηματικών διαδικασιών σε μια ποικιλία τομέων εφαρμογών. Ωστόσο, οι τυπικές τεχνικές εξόρυξης διεργασιών δεν μπορούν να κάνουν προβλέψεις για τη συμπεριφορά της διαδικασίας.

Το νευρωνικό δίκτυο Long Short-Term Memory (LSTM) είναι μια επέκταση του αναδρομικού νευρωνικού RNN, το οποίο έχει επιτύχει εξαιρετική απόδοση σε διάφορες εργασίες, ειδικά για διαδοχικά προβλήματα. Η υλοποίηση νευρωνικών δικτύων LSTM για την ανακάλυψη γεγονότων ή δραστηριοτήτων μιας επιχειρηματικής διαδικασίας μέσω της προγνωστικής ανάλυσης μπορεί να θεωρηθεί σημαντική στρατηγική ως τεχνική εξόρυξης διεργασιών και έχει χρησιμοποιηθεί με επιτυχία σε αυτόν τον τομέα .

1.2 Αντικείμενο διπλωματικής

Σε αυτή τη διπλωματική, προτείνουμε μια προσέγγιση για την ανακάλυψη των γεγονότων και δραστηριοτήτων μιας επιχειρηματικής διαδικασίας μέσω προγνωστικής ανάλυσης από ίχνη που περιέχονται σε αρχεία καταγραφής συμβάντων που λαμβάνονται από τον τραπεζικό τομέα και συγκεκριμένα από δεδομένα αιτήσεων και προσφορών δανείων. Το μοντέλο πρόβλεψης βασίζεται σε ένα αναδρομικό νευρωνικό δίκτυο LSTM που έχει εκπαιδευτεί με αρχεία καταγραφής γεγονότων, επιτρέποντας την πρόβλεψη της επόμενης δραστηριότητας σε ένα ίχνος με είσοδο μία ή περισσότερες προηγούμενες δραστηριότητες.

1.3 Οργάνωση κειμένου

Η διπλωματική εργασία αποτελείται από 7 κεφάλαια. Στο κεφάλαιο 2 παρουσιάζεται το θεωρητικό υπόβαθρο πάνω στο οποίο βασίζεται η διπλωματική. Συγκεκριμένα, παρουσιάζονται οι βασικές θεωρητικές έννοιες για την διαχείριση επιχειρησιακών διαδικασιών, την εξόρυξη διαδικασιών καθώς και η σχετική βιβλιογραφία και έρευνα για τις προβλέψεις επιχειρησιακών διαδικασιών. Στο κεφάλαιο 3 παρατίθεται το απαραίτητο

θεωρητικό υπόβαθρο για τα νευρωνικά δίκτυα μακράς βραχείας μνήμης. Στο κεφάλαιο 4 παρουσιάζεται η προτεινόμενη προσέγγιση της διπλωματικής εργασίας και συγκεκριμένα ο τρόπος προ-επεξεργασίας των δεδομένων, η αρχιτεκτονική του μοντέλου και ο τρόπος αξιολόγησης του. Στο κεφάλαιο 5 παρουσιάζεται αναλυτικά η υλοποίηση που έγινε στα πλαίσια της διπλωματικής, η πειραματική προσέγγιση, τα πειράματα, τα αποτελέσματα των προβλέψεων, τα συμπεράσματα των πειραμάτων και τέλος γίνεται σύγκριση των αποτελεσμάτων με την υπάρχουσα βιβλιογραφία. Στο κεφάλαιο 6 γίνεται μια σύνοψη της διπλωματικής μαζί με τα συνολικά συμπεράσματα που προκύπτουν συνοδευόμενα από τις προοπτικές για μελλοντική εργασία πάνω στο συγκεκριμένο αντικείμενο. Τέλος, στο κεφάλαιο 7 παρουσιάζεται η βιβλιογραφία της διπλωματικής εργασίας.

2

Θεωρητικό υπόβαθρο

2.1 Διαχείριση Επιχειρησιακών Διαδικασιών (Business Process Management)

2.1.1 Επιχειρηματική Διαδικασία

Οι ορισμοί για την έννοια της επιχειρηματικής διαδικασίας (business process) ποικίλουν και οι προσπάθειες ορισμού της επικεντρώνονται σε διάφορες οπτικές πλευρές και χαρακτηριστικά στοιχεία της. Αρχικά, ο Hammer και ο Champy (Hammer *et al.*, 1993) την διατύπωσαν ως "μια συλλογή δραστηριοτήτων που λαμβάνει ένα ή περισσότερα είδη εισόδου και δημιουργεί μια έξοδο που έχει αξία για τον πελάτη" .

Έπειτα ο Davenport ορίζει την επιχειρηματική διεργασία ως "μια συγκεκριμένη σειρά δραστηριοτήτων σε χρόνο και τόπο, με αρχή και τέλος με σαφώς καθορισμένες εισόδους και εξόδους"(Thomas H. Davenport, 1993) αναγνωρίζοντας έτσι ως ένα ακόμα σημαντικό χαρακτηριστικό του ορισμού : την σειρά περάτωσης των δραστηριοτήτων.

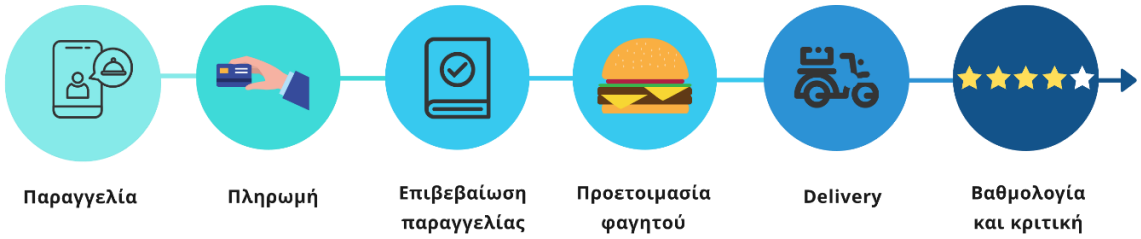
Κατά τον Mathias Weske «Μια επιχειρησιακή διαδικασία αποτελείται από ένα σύνολο δραστηριοτήτων οι οποίες εκτελούνται συντονισμένα σε ένα οργανωτικό και τεχνικό περιβάλλον. Αυτές οι δραστηριότητες από κοινού πραγματοποιούν έναν επιχειρηματικό στόχο. Κάθε επιχειρησιακή διαδικασία θεσπίζεται από μια επιχείρηση αλλά μπορεί να αλληλοεπιδρά με διαδικασίες άλλων επιχειρήσεων» (Weske, 2012).

Καταλήγουμε άρα, ότι οι επιχειρησιακές διαδικασίες αποτελούνται από εισόδους, δραστηριότητες και εξόδους. Είσοδοι είναι τα στοιχεία εκείνα που γίνονται δεκτά στη διαδικασία και που απαιτούνται για την υλοποίηση της. Αυτά μπορεί να είναι οι πρώτες ύλες, οι πληροφορίες οι εργαζόμενοι ή ακόμα και η έξοδος κάποιας άλλης διαδικασίας.

Δραστηριότητες είναι ένα από τα βήματα σε μία διαδικασία, είναι δηλαδή ένα σύνολο εργασιών που πρέπει να εκτελεστούν από ένα άτομο ή μια εφαρμογή, που έχει ένα συγκεκριμένο ρόλο (role) σε κάποια εργασία. Έξοδοι είναι οι εκροές της διαδικασίας και μπορεί να είναι η μετατροπή των πρώτων υλών σε τελικά ή ενδιάμεσα προϊόντα, οι πληροφορίες, οι εσωτερικοί ή εξωτερικοί πελάτες.

Μια επιχείρηση για να κατακτήσει με επιτυχία τους επιχειρηματικούς της στόχους χρειάζεται να υπάρχει αποτελεσματική συνεργασία των πόρων της μέσω των επιχειρησιακών διαδικασιών. Επομένως, κάθε επιχείρηση, ανεξάρτητα από το μέγεθος ή το τμήμα των δραστηριοτήτων της, είναι απαραίτητο να περιέχει επιχειρηματικές διαδικασίες. Άλλες διαδικασίες σχετίζονται με το εσωτερικό της εταιρείας και άλλες με το εξωτερικό περιβάλλον. Η αναβάθμιση και η βελτιστοποίηση των διεργασιών αυτών, οδηγεί, όχι μόνο στην επιβίωση της επιχείρησης στην αγορά, αλλά και στην διατήρηση των υψηλών επιπέδων ανταγωνιστικότητας της επιχείρησης σε σχέση με τις υπόλοιπες. Μέσω αυτών των διαδικασιών, η εταιρεία μπορεί να καλύψει τις αυξανόμενες ανάγκες των πελατών και τις μεταβαλλόμενες τάσεις της αγοράς.

Στην Εικόνα 1 φαίνεται ένα απλουστευμένο παράδειγμα επιχειρησιακής διαδικασίας για τη διαδικασία παραγγελίας φαγητού.



Εικόνα 1 Παράδειγμα επιχειρησιακής διαδικασίας

2.1.2 Ορισμός Διαχείρισης Επιχειρησιακών Διαδικασιών

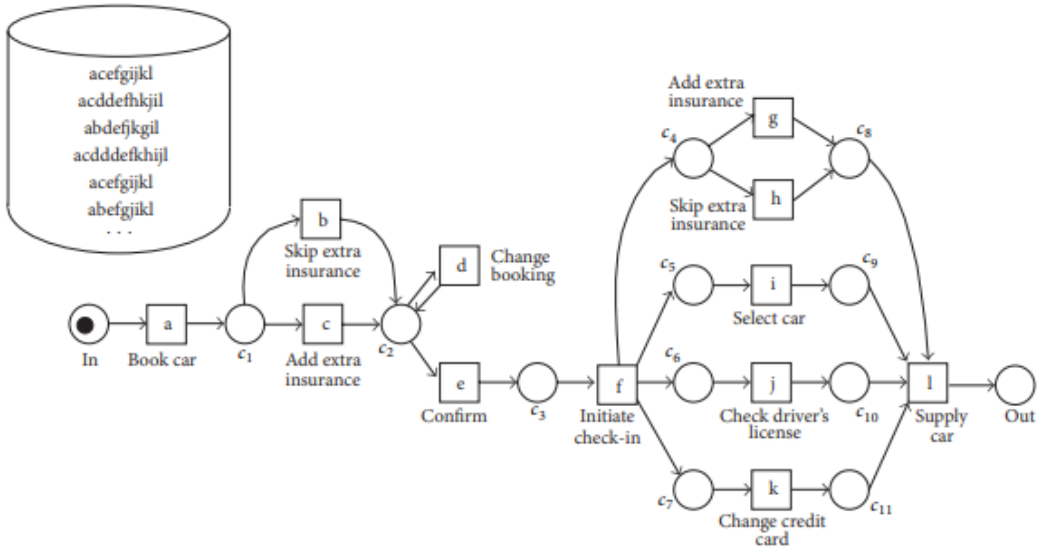
Η Διαχείριση Επιχειρησιακών Διαδικασιών (Business Process Management ή BPM) περιλαμβάνει όλες εκείνες τις έννοιες, τις μεθόδους και τις απαραίτητες τεχνικές για τον σχεδιασμό, την ανάλυση, την υλοποίηση και την διαχείριση των επιχειρησιακών διαδικασιών. Το θεμέλιο της Διαχείρισης των Επιχειρησιακών Διαδικασιών είναι η σαφής αναπαράσταση των διαδικασιών μιας επιχείρησης και των εργασιών/δραστηριοτήτων από τις οποίες αποτελούνται, μέσα σε σαφή οριοθετημένα πλαίσια. Από τη στιγμή που οι επιχειρησιακές διαδικασίες οριστούν, μπορούν να αναλυθούν, να βελτιωθούν και τελικά να εφαρμοστούν.

"Με τον όρο Διαχείριση Επιχειρησιακών Διαδικασιών – ΔΕΔ (Business Process Management-BPM) ορίζονται ένα σύνολο από έννοιες (concepts), μεθόδους (methods) και τεχνικές (techniques) για την υποστήριξη της σχεδίασης (design) και της ανάλυσης (analysis), της

διαμόρφωσης (configuration), της εκτέλεσης (enactment) και της παρακολούθησης (monitoring) επιχειρησιακών διαδικασιών."(Weske, 2012)

Η έννοια του μοντέλου διαδικασίας (process model) είναι θεμελιώδης για τη ΔΕΔ. Ένα μοντέλο διαδικασιών στοχεύει να συλλάβει τους διαφορετικούς τρόπους με τους οποίους μπορεί να διεκπερωθεί μια περίπτωση (case). Υπάρχει πληθώρα σημειογραφιών για τη μοντελοποίηση επιχειρησιακών διαδικασιών (π.χ. Petri nets, BPMN, UML και EPC). Αυτές οι σημειογραφιές έχουν το κοινό ότι οι διαδικασίες περιγράφονται ως προς τις δραστηριότητες (και πιθανώς υποδιαδικασίες). Η σειρά αυτών των δραστηριοτήτων μοντελοποιείται περιγράφοντας τις αιτιώδεις εξαρτήσεις μεταξύ τους. Επιπλέον, το μοντέλο διαδικασίας μπορεί επίσης να περιγράψει χρονικές ιδιότητες, να καθορίσει τη δημιουργία και τη χρήση δεδομένων, για παράδειγμα, για τη μοντελοποίηση αποφάσεων και να καθορίσει τον τρόπο με τον οποίο οι πόροι αλληλεπιδρούν με τη διαδικασία (π.χ. ρόλοι, κανόνες κατανομής και προτεραιότητες).

Η Εικόνα 2 δείχνει ένα μοντέλο διαδικασίας που εκφράζεται με Petri net. Το μοντέλο επιτρέπει το σενάριο $\langle a, c, e, f, g, i, j, k, l \rangle$. Αυτό είναι το σενάριο όπου γίνεται μία κράτηση αυτοκινήτου (δραστηριότητα a), προστίθεται επιπλέον ασφάλεια (δραστηριότητα c), επιβεβαιώνεται η κράτηση (δραστηριότητα e), ξεκινά η διαδικασία check-in (δραστηριότητα f), προστίθεται περισσότερη ασφάλεια (δραστηριότητα g), επιλέγεται ένα αυτοκίνητο (δραστηριότητα i), ελέγχεται η άδεια (δραστηριότητα j), χρεώνεται η πιστωτική κάρτα (δραστηριότητα k) και παρέχεται το αυτοκίνητο (δραστηριότητα l). Ένα άλλο παράδειγμα σεναρίου είναι $\langle a, c, d, d, e, f, h, k, j, i, l \rangle$ όπου η κράτηση άλλαξε δύο φορές (δραστηριότητα d) και δεν ελήφθη επιπλέον ασφάλεια κατά το check-in (δραστηριότητα h) (Van Der Aalst, 2013).



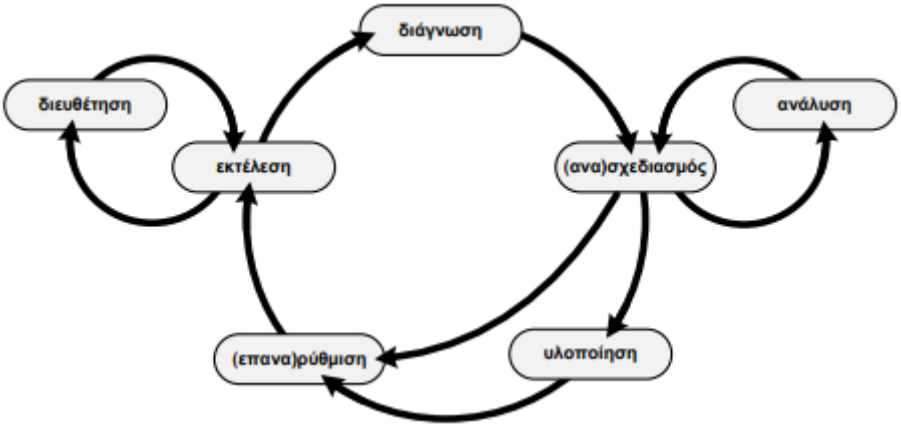
Εικόνα 2 Μοντέλο διεργασίας που εκφράζεται σε δίχτυ Petri και ένα αρχείο καταγραφής γεγονότων με ορισμένα παραδείγματα ιχνών

2.1.3 Κύκλος ζωής της Διαχείρισης Επιχειρησιακών Διαδικασιών

Ο κύκλος ζωής της ΔΕΔ που παρουσιάζεται στην Εικόνα 3 είναι το κορυφαίο μοντέλο ανάπτυξης που περιγράφει τις διάφορες φάσεις της διαχείρισης μιας επιχειρηματικής διαδικασίας. Ο κύκλος ζωής της ΔΕΔ απεικονίζει επτά φάσεις μιας επιχειρηματικής διαδικασίας και τα πληροφοριακά συστήματα που αντιστοιχούν σε αυτές. Στη φάση του (ανά)σχεδιασμού δημιουργείται ένα νέο μοντέλο διαδικασίας ή υιοθετείται ένα υπάρχον μοντέλο διαδικασίας. Στη φάση της ανάλυσης εξετάζεται το υποψήφιο μοντέλο και οι εναλλακτικές του. Μετά τη φάση του (ανά)σχεδιασμού, είτε υλοποιείται το μοντέλο (φάση υλοποίησης), ή το υπάρχον σύστημα (επανα)ρυθμίζεται (φάση επαναρύθμισης). Στη φάση της εκτέλεσης εφαρμόζεται το μοντέλο. Κατά τη διάρκεια αυτής της φάσης πραγματοποιείται και η εποπτεία της διαδικασίας. Επιπρόσθετα, είναι δυνατόν να γίνουν μικρές τροποποιήσεις στο μοντέλο χωρίς να πραγματοποιηθεί ανασχεδιασμός (φάση διευθέτησης). Στη φάση της διάγνωσης, αναλύεται η διαδικασία που εφαρμόστηκε, ενώ τα αποτελέσματα της διάγνωσης μπορούν να πυροδοτήσουν μία νέα φάση ανασχεδιασμού της διαδικασίας. Η εξόρυξη διαδικασιών είναι ένα πολύτιμο εργαλείο για τις περισσότερες από τις φάσεις της Εικόνα 3 .

Η εξόρυξη διεργασιών είναι ένας ειδικός τύπος εξόρυξης δεδομένων, που επικεντρώνεται συγκεκριμένα στην ανάλυση ιστορικών δεδομένων των εκτελέσεων διαδικασίας με τη μορφή αρχείων καταγραφής γεγονότων. Οι τεχνικές εξόρυξης διεργασιών είναι σε θέση να παρέχουν πληροφορίες για την τρέχουσα εκτέλεση της επιχειρηματικής διαδικασίας με βάση τα παρατηρούμενα γεγονότα όπως καταγράφονται στο αρχείο καταγραφής γεγονότων. Περισσότερα σχετικά με την εξόρυξη διεργασιών παρουσιάζονται στην επόμενη ενότητα.

Προφανώς, η φάση της διάγνωσης μπορεί να επωφεληθεί από την εξόρυξη διαδικασιών. Ωστόσο, η συμβολή της εξόρυξης διαδικασιών δεν περιορίζεται σε αυτή τη φάση (διάγνωση). Για παράδειγμα, οι τεχνικές εξόρυξης διαδικασιών μπορούν να χρησιμοποιηθούν στη φάση της εκτέλεσης για λειτουργική υποστήριξη (operational support). Οι προβλέψεις και οι συστάσεις που βασίζονται σε μοντέλα που έχουν προκύψει από ιστορικά στοιχεία μπορούν να χρησιμοποιηθούν προς όφελος διαδικασιών που βρίσκονται σε εξέλιξη. Παρόμοιες δομές υποστήριξης αποφάσεων μπορούν να χρησιμοποιηθούν για να εισάγουν τροποποιήσεις στις διαδικασίες και για να καθοδηγήσουν τη διαδικασία επαναρύθμισης τους.(Van Der Aalst *et al.*, 2012)



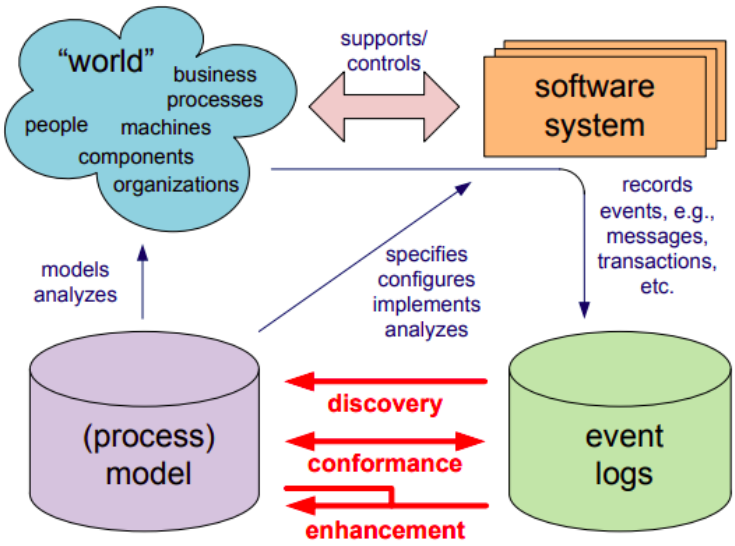
Εικόνα 3 : Ο κύκλος ζωής της διοίκησης επιχειρησιακών διαδικασιών

2.2 Εξόρυξη Διεργασιών (Process Mining)

Η εξόρυξη διεργασιών στοχεύει στην ανακάλυψη, την παρακολούθηση και τη βελτίωση πραγματικών διαδικασιών εξάγοντας γνώσεις από αρχεία καταγραφής γεγονότων που είναι άμεσα διαθέσιμα στα σημερινά συστήματα πληροφοριών (Van Der Aalst, 2012). Κατά την τελευταία δεκαετία υπήρξε μια θεαματική αύξηση των δεδομένων γεγονότων και οι τεχνικές εξόρυξης διεργασιών έχουν ωριμάσει σημαντικά. Ως αποτέλεσμα, οι τάσεις στη διοίκηση που σχετίζονται με τη βελτίωση των διαδικασιών των επιχειρήσεων και τη συμμόρφωση με διάφορα πρότυπα μπορούν πλέον να επωφεληθούν από την εξόρυξη διαδικασιών. Αφετηρία για την εξόρυξη διαδικασιών είναι το αρχείο καταγραφής γεγονότων (event log). Κάθε γεγονός σε ένα τέτοιο αρχείο καταγραφής αναφέρεται σε μια δραστηριότητα (activity) (δηλαδή, ένα καλά καθορισμένο βήμα σε κάποια διαδικασία) και σχετίζεται με μια συγκεκριμένη περίπτωση (case).

2.2.1 Τύποι Εξόρυξης Διαδικασιών

Τα γεγονότα που ανήκουν σε μια περίπτωση διατάσσονται και μπορούν να θεωρηθούν ως ένα «τρέξιμο» της διαδικασίας. Τα αρχεία καταγραφής γεγονότων ενδέχεται να αποθηκεύουν πρόσθετες πληροφορίες σχετικά με τα γεγονότα (events). Στην πραγματικότητα, όποτε είναι δυνατόν, οι τεχνικές εξόρυξης διεργασιών χρησιμοποιούν επιπλέον πληροφορίες όπως οι πόροι (δηλαδή, άτομο ή συσκευή) που εκτελεί ή ξεκινά τη δραστηριότητα, τη χρονική σήμανση του γεγονότος ή τα στοιχεία δεδομένων που καταγράφονται με το συμβάν (π.χ., το μέγεθος μιας παραγγελίας). Τα αρχεία καταγραφής γεγονότων μπορούν να χρησιμοποιηθούν για τη διεξαγωγή τριών τύπων εξόρυξης διεργασιών, όπως φαίνεται στην Εικόνα 4 (Van Der Aalst *et al.*, 2012).



Εικόνα 4 Θέση των τριών κύριων τύπων εξόρυξης διεργασιών: (α) ανακάλυψη, (β) έλεγχος συμμόρφωσης και (γ) βελτίωση

Ο πρώτος τύπος εξόρυξης διαδικασιών είναι η ανακάλυψη. Μια τεχνική ανακάλυψης λαμβάνει ένα αρχείο καταγραφής γεγονότων και παράγει ένα μοντέλο χωρίς τη χρήση εκ των προτέρων πληροφοριών. Η ανακάλυψη διεργασιών είναι η πιο εξέχουσα τεχνική εξόρυξης διεργασιών. Για πολλούς οργανισμούς είναι εκπληκτικό να βλέπουμε ότι οι υπάρχουσες τεχνικές είναι πράγματι σε θέση να ανακαλύψουν πραγματικές διαδικασίες μόνο με βάση παραδείγματα συμπεριφορών που αποθηκεύονται σε αρχεία καταγραφής γεγονότων.

Ο δεύτερος τύπος εξόρυξης διεργασιών είναι η συμμόρφωση. Εδώ, ένα υπάρχον μοντέλο διαδικασίας συγκρίνεται με ένα αρχείο καταγραφής γεγονότων της ίδιας διαδικασίας. Ο έλεγχος συμμόρφωσης μπορεί να χρησιμοποιηθεί για να ελέγξει εάν η πραγματικότητα, όπως καταγράφεται στο αρχείο καταγραφής, συμμορφώνεται με το μοντέλο και αντίστροφα.

Ο τρίτος τύπος εξόρυξης διεργασιών είναι η βελτίωση. Εδώ, η ιδέα είναι να επεκταθεί ή να βελτιωθεί ένα υπάρχον μοντέλο διεργασίας χρησιμοποιώντας έτσι πληροφορίες σχετικά με την πραγματική διαδικασία που καταγράφεται σε κάποιο αρχείο καταγραφής γεγονότων. Ενώ ο έλεγχος της συμμόρφωσης μετρά την ευθυγράμμιση μεταξύ μοντέλου και πραγματικότητας, αυτός ο τρίτος τύπος διαδικασίας εξόρυξης στοχεύει στην αλλαγή ή επέκταση του a priori μοντέλου. Για παράδειγμα, χρησιμοποιώντας τις χρονικές σημάνσεις στο αρχείο καταγραφής γεγονότων, μπορεί κάποιος να επεκτείνει το μοντέλο για να εμφανίσει σημεία συμφόρησης, επίπεδα υπηρεσίας και χρόνους απόδοσης.

2.2.2 Αρχεία Καταγραφής Γεγονότων (Event Logs)

Τα δεδομένα ψηφιακών γεγονότων είναι παντού - σε κάθε τομέα, σε κάθε οικονομία, σε κάθε οργανισμό και σε κάθε σπίτι - και θα συνεχίσουν να αυξάνονται εκθετικά (Manyika *et al.*, 2011). Η πανταχού παρουσία τέτοιων δεδομένων επιτρέπει νέες μορφές ανάλυσης της διαδικασίας, δηλαδή βασίζεται σε παρατηρούμενα γεγονότα και όχι σε χειροποίητα μοντέλα. Αφετηρία για τη διαδικασία της εξόρυξης είναι ένα αρχείο καταγραφής γεγονότων (event log).

Τυπικά, ένα γεγονός e είναι μια πλειάδα e = (a, c, τ, D) όπου: a ∈ A είναι το όνομα της δραστηριότητας της διεργασίας που σχετίζεται με το γεγονός e (δηλ. ποιες εργασίες εκτελέστηκαν), c ∈ C είναι το "case-id", που είναι ένα αναγνωριστικό της συγκεκριμένης περίπτωσης της επιχειρηματικής διεργασίας, τ ∈ T είναι μια χρονική σήμανση που δείχνει τον χρόνο εκτέλεσης της συγκεκριμένης δραστηριότητας, $D \equiv \{(d_1, v_1), \ldots, (d_m, v_m)\}$ είναι ένα σύνολο ζευγών χαρακτηριστικών που σχετίζονται με την εκτέλεση του γεγονότος e. Υποθέτουμε την παρουσία ενός χειριστή προβολής π, ο οποίος επιτρέπει την εξαγωγή συγκεκριμένων χαρακτηριστικών από ένα γεγονός. Συγκεκριμένα, δεδομένου ενός γεγονότος e = (a, c, τ, D), ορίζουμε $\pi_A(e) = a, \pi_C(e) = c, \pi_T(e) = \tau$, και $\pi_{di}(e) = v_i$. Το σύνολο γεγονότων Ɛ δείχνει το σύνολο όλων των πιθανών γεγονότων. Επιπλέον, αναφερόμαστε ως Ɛ ∗ το σύνολο όλων των πιθανών ακολουθιών πάνω από το Ɛ. Ένα ίχνος t = <e1,. . . , e_n> ∈ Ɛ ∗, μήκους | t | = n, είναι οποιαδήποτε από αυτές τις ακολουθίες, όπου ∀ 1 ≤ i ≤ | t |, $\pi_C(e_i)$ = c.

Ένα αρχείο καταγραφής γεγονότων σε ένα σύνολο γεγονότων Ɛ είναι ένα σύνολο ιχνών L ⊆Ɛ∗ έτσι ώστε κάθε γεγονός να εμφανίζεται το πολύ μία φορά σε ολόκληρο το αρχείο καταγραφής, δηλαδή, για οποιοδήποτε t1, t2 ∈ L, t1 ≠ t2: set (t1) ∩set (t2) = ∅, όπου το set (t) μετατρέπει μια ακολουθία t σε ένα σύνολο με τα ίδια στοιχεία (Navarin *et al.*, 2018).

Ένα παράδειγμα event log παρουσιάζεται πιο κάτω στον Πίνακας 1:

Πίνακας 1 Παράδειγμα αρχείου καταγραφής γεγονότων (Event Log)

Case Id	Event Id	Activity	Timestamp	Resource
5781	2	X-ray	29-05-2021 18:53:00	Dr. Dionysis
5782	5	Blood test	29-05-2021 20:33:00	Dr. Despina
5781	1	Blood test	29-05-2021 12:13:00	Dr. Maria
5781	3	CT scan	29-05-2021 19:12:00	Dr. Mattheos
5781	4	Surgery	29-05-2021 20:30:00	Dr. Marios
5782	6	Payment	29-05-2021 20:38:00	Dr. Tasoula
5783	7	Check in	29-05-2021 20:42:00	Dr. Christos
5783	8	Blood Test	29-05-2021 20:48:00	Dr. Dimitris

Από τον πίνακα αυτό μπορούμε να συμπεράνουμε ότι οι δραστηριότητες της διεργασίας που καταγράφονται στο κομμάτι του event log που βλέπουμε είναι οι:

a) X-ray

b) Blood Test

c) CT scan

d) Surgery

e) Payment

f) Check in

Θεωρώντας ως ίχνος μια ακολουθία από αυτές τις δραστηριότητες, κάθε περίπτωση που ανήκει στο event log αντιστοιχεί σε ένα ίχνος όπως παρουσιάζεται στον επόμενο πίνακα (Πίνακας 2):

Πίνακας 2 Αντιστοίχιση Case και Trace στο παράδειγμα

Case ID	Trace
5781	<a,b,c,d>
5782	<b,e>
5783	<f,b>

2.2.3 Εξόρυξη Διαδικασιών στη Python

Το Pm4py παρέχει ένα λογισμικό εξόρυξης διεργασιών που είναι εύκολα επεκτάσιμο, επιτρέπει αλγοριθμική προσαρμογή και επιτρέπει στον χρήστη να διεξάγει εύκολα πειράματα μεγάλης κλίμακας (Berti, van Zelst and van der Aalst, 2019). Ο κόσμος της επιστήμης δεδομένων, τόσο στην κλασική επιστήμη των δεδομένων όσο και στη μηχανική μάθηση, χρησιμοποιεί σε μεγάλο βαθμό την Python. Άλλες βιβλιοθήκες, αν και με μικρότερο αριθμό χαρακτηριστικών, υπάρχουν ήδη για τη γλώσσα Python. Η βιβλιοθήκη bupaR υποστηρίζει διαδικασία εξόρυξης στη στατιστική γλώσσα R, η οποία χρησιμοποιείται ευρέως στην επιστήμη των δεδομένων.

Τα κύρια σημεία εστίασης της νέας βιβλιοθήκης PM4Py την οποία θα χρησιμοποιήσω στη συνέχεια στην παρούσα διπλωματική εργασία είναι:

• Μείωση του φραγμού για την αλγοριθμική ανάπτυξη και προσαρμογή κατά την εκτέλεση μιας ανάλυσης εξόρυξης διεργασιών σε σύγκριση με τα υπάρχοντα ακαδημαϊκά εργαλεία όπως τα ProM, RAPIdProM και Apromore.

• Επιτρέπει την εύκολη ενσωμάτωση αλγορίθμων εξόρυξης διεργασιών με αλγόριθμους από άλλα πεδία επιστήμης δεδομένων, που εφαρμόζονται σε διάφορα υπερσύγχρονα πακέτα Python.

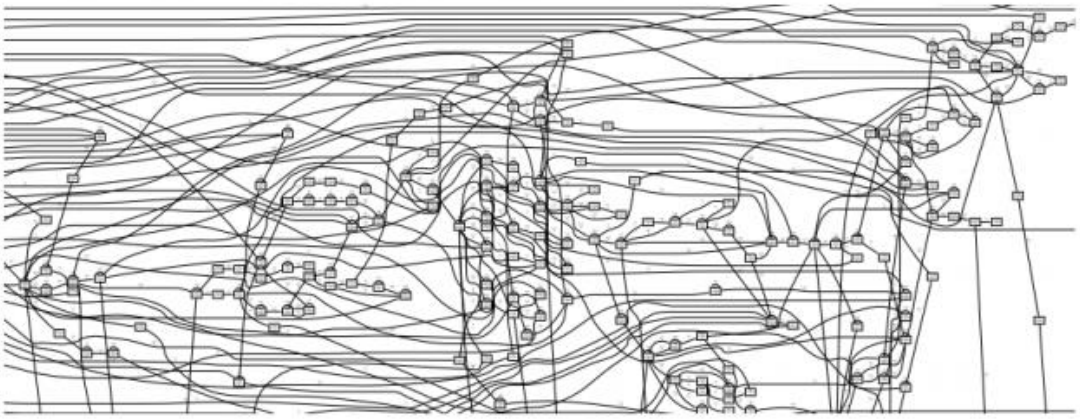
• Δημιουργεί ένα συνεργατικό οικοσύστημα που επιτρέπει στους ερευνητές και τους επαγγελματίες να μοιράζονται πολύτιμο κώδικα και αποτελέσματα με τον κόσμο της εξόρυξη διαδικασιών.

• Παρέχει υποστήριξη χρηστών μέσω ενός πλούσιου υλικού τεκμηρίωσης σχετικά με τις τεχνικές εξόρυξης διεργασιών που διατίθενται στη βιβλιοθήκη.

• Αλγοριθμική σταθερότητα μέσω αυστηρών δοκιμών.

2.2.4 Μοντέλα Διεργασιών

Τα αρχεία καταγραφής γεγονότων που δημιουργούνται από τα συστήματα πληροφοριών που χρησιμοποιούνται σε κάθε οργανισμό είναι η είσοδος για έναν αλγόριθμο ανακάλυψης διεργασιών. Ο αλγόριθμος ανακάλυψης διεργασίας δημιουργεί ένα μοντέλο διεργασίας που είναι η διαδικασία που ακολουθείται επί του παρόντος στον οργανισμό. Στη φάση βελτιστοποίησης της διαδικασίας, αρκετοί αλγόριθμοι μπορούν να εφαρμοστούν στη διαδικασία προκειμένου να προταθεί αυτόματα μια βελτιωμένη έκδοση. Η Εικόνα 5 δείχνει ένα παράδειγμα μοντέλου διεργασίας που μοιάζει με σπαγγέτι. Τα λεγόμενα “μοντέλα σπαγγέτι” (spaghetti-like models), αποδεικνύουν ότι τα παραγόμενα μοντέλα είναι πολλές φορές δύσκολο να γίνουν κατανοητά, υστερούν από θέμα εκφραστικότητας και είναι αδύνατο να ερμηνευθούν με ανθρώπινο μάτι. Λίγοι αλγόριθμοι βελτιστοποίησης είναι σε θέση να λειτουργήσουν με αυτούς τους τύπους διαδικασιών (Chinces and Salomie, 2015).



Εικόνα 5 Παράδειγμα ενός "spaghetti like" μοντέλου

Τα μοντέλα διεργασιών δημιουργήθηκαν για να συντάσσουν και να εξηγούν τη δρομολόγηση των περιπτώσεων των επιχειρηματικών διαδικασιών. Υπάρχουν πολλές και διαφορετικές γλώσσες μοντελοποίησης διαδικασιών. Οι περισσότερες από αυτές, παρέχουν μια γραφική σημειογραφία με κόμβους διάσπασης και με κόμβους ένωσης. Για την αναπαράσταση ενός μοντέλου διεργασιών μπορούν να χρησιμοποιηθούν διάφορες σημειογραφίες όπως Petri nets, BPMN, διαγράμματα UML κ.ά.

2.3 Προβλέψεις Επιχειρηματικών Διαδικασιών

Η ανάπτυξη τεχνολογικών λύσεων για ανάλυση event logs για την ανακάλυψη επιχειρηματικών διαδικασιών χρησιμοποιώντας τις αρχές της εξόρυξης δεδομένων έχει μελετηθεί προηγουμένως στα (van der Aalst, 2016), (Baier, Mendling and Weske, 2014) .Οι πιο σχετικές προτάσεις που σχετίζονται με την προσέγγιση που προτείνεται στη διπλωματική θα συζητηθούν σε αυτήν την ενότητα. Παρόλο που παλαιότερες τεχνικές δεν είναι σε θέση να προβλέψουν σε πραγματικό χρόνο τις επόμενες δραστηριότητες που πρόκειται να εκτελεστούν σε μια επιχειρηματική διαδικασία, οι τεχνικές που βασίζονται σε νευρωνικά δίκτυα LSTM, όπως προτείνεται και στη διπλωματική αυτή, μπορούν να βοηθήσουν στην ανακάλυψη μοντέλων επιχειρηματικών διαδικασιών.

Υπάρχουν μερικές προσεγγίσεις που χρησιμοποιούν μοτίβα και στατιστικά μοντέλα για την πρόβλεψη δραστηριοτήτων σε επιχειρηματικές διαδικασίες. Η προσέγγιση που περιγράφεται στο (Ceci *et al.*, 2014), στοχεύει στον εντοπισμό μερικών μοντέλων επιχειρηματικής διαδικασίας που θα χρησιμοποιηθούν για την εκπαίδευση προγνωστικών μοντέλων. Χρησιμοποιεί την εξόρυξη ακολουθιών για τον εντοπισμό συχνών προθεμάτων των ιχνών. Για κάθε πρόθεμα, ένα μοντέλο παλινδρόμησης εκπαιδεύεται για να προβλέψει τον υπολειπόμενο χρόνο έως την ολοκλήρωση και ένα δέντρο αποφάσεων εκπαιδεύεται για να προβλέψει το επόμενο συμβάν. Ο αλγόριθμος προσδιορίζει το κατάλληλο πρόθεμα της περίπτωσης που τρέχει εκείνη την ώρα για να επιλέξει το μοντέλο παλινδρόμησης και δέντρο απόφασης για την πρόβλεψη του υπολειπόμενου χρόνου ολοκλήρωσης και του επόμενου συμβάντος. Η πειραματική αξιολόγηση χρησιμοποιεί δύο σύνολα δεδομένων, αποδίδοντας τιμές ακρίβειας πρόβλεψης για το επόμενο συμβάν περίπου 65% σε ένα σύνολο δεδομένων και περίπου 50% στο άλλο, ανάλογα με τον τύπο του δέντρου αποφάσεων που χρησιμοποιείται και το όριο συχνότητας για τα προθέματα.

Η προσέγγιση που περιγράφεται από τους (Lakshmanan *et al.*, 2015) αποτελείται από πέντε βήματα. Πρώτον, ένα μοντέλο διεργασίας εξορύσσεται από υπάρχοντα αρχεία καταγραφής. Για κάθε διαχωρισμό XOR στο μοντέλο, ένα δέντρο αποφάσεων εξορύσσεται από τα δεδομένα των περιπτώσεων. Αυτά τα δέντρα χρησιμοποιούνται στη συνέχεια για τον υπολογισμό των πιθανοτήτων μετάβασης κατάστασης για ένα HMM που είναι ειδικό για την υπόθεση που θα προβλεφθεί. Αυτό το HMM χρησιμοποιείται στη συνέχεια για να προβλέψει τις πιθανότητες του ακόλουθου συμβάντος. Η προσέγγιση αξιολογείται σε προσομοιωμένα δεδομένα. Αφού εκπαιδευτούν τα δέντρα αποφάσεων για το HMM στο μισό από τα ίχνη (σετ προπόνησης), κάθε ίχνος στο σετ δοκιμής κόβεται σε πρόθεμα και επίθεμα σε τυχαίο σημείο. Για κάθε πρόθεμα, αναφέρεται η πιθανότητα καταγραφής του αντίστοιχου επιθέματος.

Η χρήση νευρωνικών δικτύων για την πρόβλεψη δραστηριοτήτων είναι ένα πρόσφατο πεδίο. Στο (Evermann, Rehse and Fettke, 2016), ο συγγραφέας παρουσιάζει μια προσέγγιση για την πρόβλεψη του επόμενου συμβάντος της διαδικασίας χρησιμοποιώντας βαθιά μάθηση με βάση ένα επαναλαμβανόμενο νευρωνικό δίκτυο LSTM. Η προτεινόμενη προσέγγιση βασίζεται στην πρόβλεψη της επόμενης λέξης σε μια πρόταση (επεξεργασία φυσικής γλώσσας), δηλαδή, ερμηνεύοντας τα αρχεία καταγραφής της διαδικασίας ως κείμενο, τα ίχνη της διαδικασίας ως προτάσεις και τα γεγονότα επεξεργασίας ως λέξεις. Η υλοποίηση του δικτύου LSTM γίνεται μέσω μιας αρχιτεκτονικής δύο κρυφών επιπέδων. Η προσέγγιση αξιολογείται σε δύο πραγματικά σύνολα δεδομένων που χρησιμοποιούνται συνήθως στο state-of-the-art, παρουσιάζοντας καλύτερα αποτελέσματα στην ακρίβεια των προβλέψεων.

Ομοίως, στο (Tax *et al.,* 2017), οι συγγραφείς προτείνουν μια μέθοδο βασισμένη στο LSTM που επιτρέπει την πρόβλεψη της επόμενης δραστηριότητας και της χρονικής διάρκειας της ανά περίπτωση που περιέχεται σε ένα αρχείο γεγονότος. Οι συγγραφείς αναφέρουν ότι τα αποτελέσματα του νευρωνικού δικτύου είναι εξαιρετικά χρησιμοποιώντας σύνολα δεδομένων πραγματικής ζωής σε σύγκριση με τις παραδοσιακές μεθόδους αυτόματης μάθησης. Επιπλέον, οι συγγραφείς καταλήγουν στο συμπέρασμα ότι η πρόβλεψη της επόμενης δραστηριότητας και της χρονικής σήμανσής της μέσω ενός μόνο μοντέλου αποδίδει μεγαλύτερη ακρίβεια από την πρόβλεψη της χρήσης ξεχωριστών μοντέλων(Tello-Leal *et al.*, 2018).

Ακόμα στο (Evermann, Rehse and Fettke, 2016) εφαρμόζουν επίσης δίκτυα LSTM για να προβλέψουν τον τύπο του επόμενου γεγονότος ενός case. Σε αντίθεση με το (Tax *et al.*, 2017), αυτή η προσέγγιση χρησιμοποιεί την ενσωματωμένη διάσταση (embedded dimension) των LSTM για να μειώσει το μέγεθος της εισόδου και να συμπεριλάβει πρόσθετα χαρακτηριστικά όπως οι πόροι που σχετίζονται με κάθε συμβάν. Η αρχιτεκτονική του δικτύου περιλαμβάνει δύο κρυμμένα επίπεδα LSTM.

Στο (Tello-Leal *et al.*, 2018) παρουσιάστηκε ένα δίκτυο LSTM που εφαρμόζεται σε αρχεία καταγραφής γεγονότων από τον τομέα Internet of Things (IoT) στο πλαίσιο της βιομηχανίας 4.0. Σε αυτό το έργο, έχει προταθεί μια μεθοδολογία για την πρόβλεψη των επόμενων δραστηριοτήτων μιας επιχειρηματικής διαδικασίας. Η μεθοδολογία βασίζεται σε ένα επαναλαμβανόμενο νευρωνικό δίκτυο LSTM και εκμεταλλεύεται αρχεία καταγραφής γεγονότων για να κάνει προβλέψεις σχετικά με την εκτέλεση των περιπτώσεων. Η προτεινόμενη μεθοδολογία περιλαμβάνει προ-επεξεργασία, την κατηγοριοποίηση και το μοντέλο πρόβλεψης καταγραφής γεγονότων, και επιτρέπει τον προσδιορισμό των φάσεων που απαιτούνται για την πρόβλεψη της επόμενης δραστηριότητας ή συμβάντος, μέσω της εφαρμογής του νευρωνικού δικτύου LSTM. Η δοκιμή που πραγματοποιήθηκε στο εκπαιδευμένο δίκτυο LSTM δείχνει ότι έχει την ικανότητα να προβλέψει την επόμενη δραστηριότητα ενός μοντέλου επιχειρηματικής διαδικασίας.

3

Νευρωνικά Δίκτυα Μακράς Βραχείας Μνήμης

3.1 Τεχνητά Νευρωνικά Δίκτυα

Εμπνευσμένα από τα βιολογικά νευρωνικά δίκτυα, τα τεχνητά νευρωνικά δίκτυα (artificial neural networks- ANN) είναι παράλληλα υπολογιστικά συστήματα που αποτελούνται από έναν εξαιρετικά μεγάλο αριθμό απλών επεξεργαστών με πολλές διασυνδέσεις. Τα μοντέλα ANN επιχειρούν να χρησιμοποιήσουν κάποιες «οργανωτικές» αρχές που πιστεύεται ότι χρησιμοποιούνται στον άνθρωπο. Ένας τύπος δικτύου βλέπει τους κόμβους ως «τεχνητούς νευρώνες». Αυτά ονομάζονται τεχνητά νευρικά δίκτυα (ANN). Ένας τεχνητός νευρώνας είναι ένα υπολογιστικό μοντέλο εμπνευσμένο από τους φυσικούς νευρώνες. Δεδομένου ότι η λειτουργία των ANN είναι η επεξεργασία πληροφοριών, χρησιμοποιούνται κυρίως σε πεδία που σχετίζονται με αυτήν.

Υπάρχει μια μεγάλη ποικιλία ANN που χρησιμοποιούνται για τη μοντελοποίηση πραγματικών νευρωνικών δικτύων και μελετούν τη συμπεριφορά και τον έλεγχο σε ζώα και μηχανήματα, αλλά υπάρχουν και ANN που χρησιμοποιούνται για μηχανικούς σκοπούς, όπως αναγνώριση προτύπων, πρόβλεψη και συμπίεση δεδομένων. Αυτά βασικά αποτελούνται από εισόδους (όπως οι συνάψεις στον άνθρωπο), οι οποίες πολλαπλασιάζονται με βάρη. Τα βάρη που εκχωρούνται με κάθε βέλος αντιπροσωπεύουν τη ροή πληροφοριών. Αυτά τα βάρη υπολογίζονται στη συνέχεια από μια μαθηματική συνάρτηση που καθορίζει την ενεργοποίηση του νευρώνα. Μια άλλη συνάρτηση υπολογίζει την έξοδο του τεχνητού νευρώνα (μερικές φορές σε εξάρτηση από ένα ορισμένο κατώφλι). Οι νευρώνες αυτού του δικτύου αθροίζουν τις εισροές τους. Δεδομένου ότι οι νευρώνες εισόδου έχουν μόνο μία είσοδο, η έξοδος τους θα είναι η είσοδος που έλαβαν πολλαπλασιασμένη επί το βάρος.

Εάν το βάρος είναι υψηλό, τότε η είσοδος θα είναι ισχυρή. Ρυθμίζοντας τα βάρη ενός τεχνητού νευρώνα μπορούμε να αποκτήσουμε την έξοδο που θέλουμε για συγκεκριμένες εισόδους. Αλλά

όταν έχουμε ένα ANN εκατοντάδων ή χιλιάδων νευρώνων, θα ήταν πολύ περίπλοκο να βρούμε με το χέρι όλα τα απαραίτητα βάρη. Μπορούμε όμως να βρούμε αλγόριθμους που μπορούν να προσαρμόσουν τα βάρη του ANN προκειμένου να επιτύχουν την επιθυμητή έξοδο από το δίκτυο. Αυτή η διαδικασία προσαρμογής των βαρών ονομάζεται μάθηση ή εκπαίδευση. Η εκπαίδευση ξεκινά με τυχαίο σφάλμα το οποίο θα είναι ελάχιστο (Gupta, 2013).

3.2 Αναδρομικά νευρωνικά δίκτυα (RNN)

Ένα νευρωνικό δίκτυο αποτελείται από ένα στρώμα μονάδων εισόδου, ένα στρώμα μονάδων εξόδου και πολλαπλά επίπεδα μεταξύ τους τα οποία αναφέρονται ως κρυφές μονάδες. Οι έξοδοι των μονάδων εισόδου σχηματίζουν τις εισόδους των μονάδων του πρώτου κρυφού επιπέδου (δηλαδή, το πρώτο στρώμα των κρυφών μονάδων), και οι έξοδοι των μονάδων κάθε κρυμμένου επιπέδου σχηματίζουν την είσοδο για κάθε επόμενο κρυφό επίπεδο. Οι έξοδοι του τελευταίου κρυφού στρώματος αποτελούν την είσοδο για το επίπεδο εξόδου. Η έξοδος κάθε μονάδας είναι μια συνάρτηση πάνω από το σταθμισμένο άθροισμα των εισόδων της. Τα βάρη αυτού του σταθμισμένου αθροίσματος που εκτελούνται σε κάθε μονάδα μαθαίνονται μέσω βελτιστοποίησης βάσει κλίσης (gradient-based optimization) από δεδομένα εκπαίδευσης που αποτελούνται από παραδείγματα εισόδων και επιθυμητές εξόδους για αυτά τα παραδείγματα εισόδων.

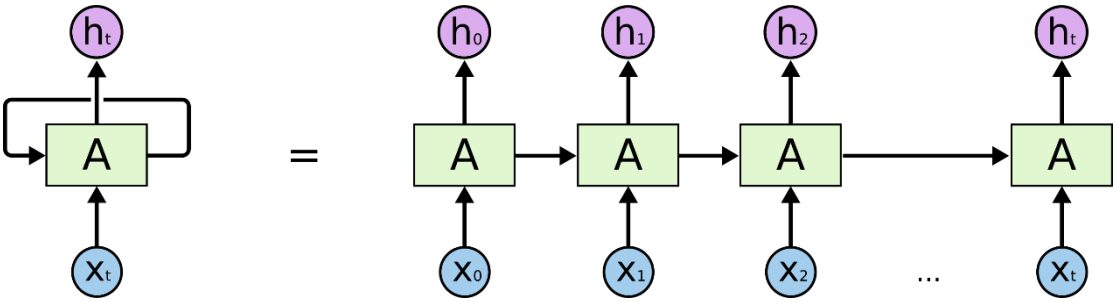
Τα αναδρομικά νευρωνικά δίκτυα (RNN) είναι ένας ειδικός τύπος νευρωνικών δικτύων όπου οι συνδέσεις μεταξύ νευρώνων σχηματίζουν έναν κατευθυνόμενο κύκλο. Τα RNN μπορούν να ξεδιπλωθούν, όπως φαίνεται στην εικόνα Εικόνα 6. Κάθε βήμα στο ξεδίπλωμα αναφέρεται ως χρονικό βήμα, όπου x_t είναι η είσοδος στο χρονικό βήμα t. Τα RNN μπορούν να πάρουν μια αυθαίρετη ακολουθία μήκους ως είσοδο, παρέχοντας στο RNN μια αναπαράσταση χαρακτηριστικών ενός στοιχείου της ακολουθίας σε κάθε χρονικό βήμα. Το s_t είναι η κρυφή κατάσταση στο βήμα t και περιέχει πληροφορίες που εξάγονται από όλα τα χρονικά βήματα έως το t. Η κρυφή κατάσταση s ενημερώνεται με πληροφορίες για τη νέα είσοδο x_t μετά από κάθε βήμα όπως φαίνεται στην εξίσωση 1:

$$s_t = f(Ux_t + Ws_{t-1}) \quad (1)$$

τα U και W είναι διανύσματα βαρών πάνω από τις νέες εισόδους και την κρυφή κατάσταση αντίστοιχα. Η συνάρτηση f, γνωστή ως συνάρτηση ενεργοποίησης, είναι συνήθως είτε η υπερβολική εφαπτομένη είτε η λογιστική συνάρτηση, που συχνά αναφέρεται ως η σιγμοειδής συνάρτηση:

$$sigmoid(x) = \frac{1}{1+e^{(-x)}} \quad (2)$$

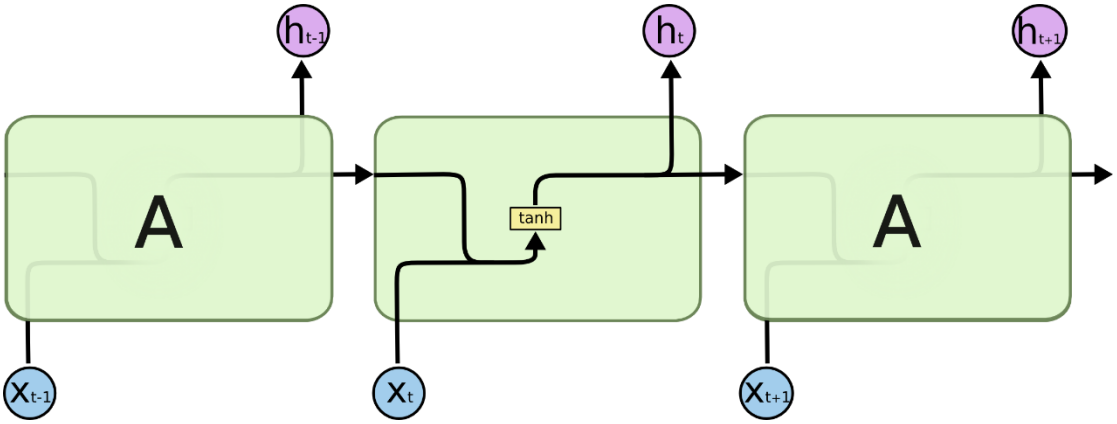
Στη βιβλιογραφία του νευρωνικού δικτύου, η σιγμοειδής συνάρτηση αντιπροσωπεύεται συχνά με το γράμμα σ. (Tax *et al.*, 2017)



Εικόνα 6 Ένα απλό RNN

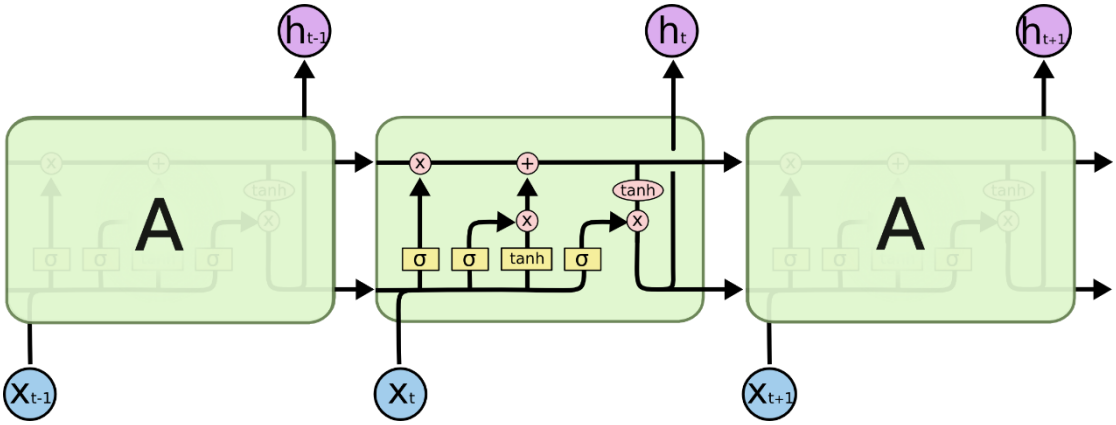
3.3 Νευρωνικό Δίκτυο Μακράς-Βραχείας Μνήμης – LSTM

Το νευρωνικό δίκτυο Μακράς-Βραχείας Μνήμης (LSTM) είναι μια ειδική αρχιτεκτονική Recurrent Neural Network που προτάθηκε για πρώτη φορά από τους Hochreiter και Schmidhuber το 1997 που έχει ισχυρές δυνατότητες μοντελοποίησης για μακροπρόθεσμες εξαρτήσεις. Όπως όλα τα RNN, έτσι και τα LSTM έχουν τη μορφή μιας αλυσίδας επαναλαμβανόμενων μονάδων από νευρωνικά δίκτυα με τη διαφορά όμως ότι η δομή αυτών των μονάδων τους είναι αρκετά πιο περίπλοκή από αυτή των απλών RNN η οποία αποτελείται από ένα επίπεδο με μία tanh συνάρτηση ενεργοποίησης όπως φαίνεται στην Εικόνα 7.



Εικόνα 7 Η επαναλαμβανόμενη μονάδα σε ένα τυπικό RNN περιέχει ένα μόνο στρώμα.

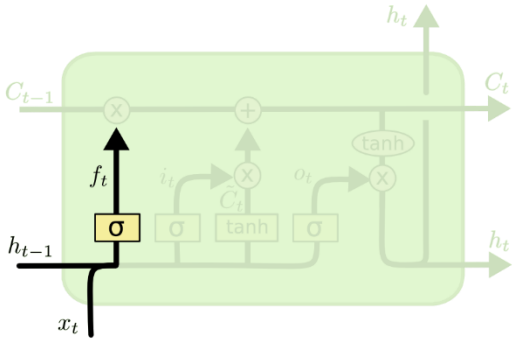
Τα LSTM είναι επίσης σχεδιασμένα με μια παρόμοια αλυσίδα, όμως η μονάδα που επαναλαμβάνεται έχει διαφορετική δομή. Περιλαμβάνει αντί για ένα επίπεδο νευρωνικού δικτύου, τέσσερα που αλληλοεπιδρούν μεταξύ τους όπως φαίνεται στην Εικόνα 8.



Εικόνα 8 Η επαναλαμβανόμενη μονάδα σε ένα LSTM περιέχει τέσσερα επίπεδα αλληλεπίδρασης

Το κλειδί για τη λειτουργία του LSTM είναι η κατάσταση του κελιού μνήμης. Η κατάσταση του κελιού C_t είναι σαν μια ζώνη που διατρέχει την αλυσίδα. Η γραμμή αυτή επεκτείνεται στην αλυσίδα με μικρές γραμμικές αλλαγές στο περιεχόμενο της. Το LSTM έχει την δυνατότητα να αφαιρεί ή να προσθέτει πληροφορία στο κύτταρο κατάστασης, με την χρήση δομών που λέγονται πύλες (gates) που χρησιμοποιούν σιγμοειδείς συναρτήσεις ενεργοποίησης με τιμές στο εύρος [0,1].

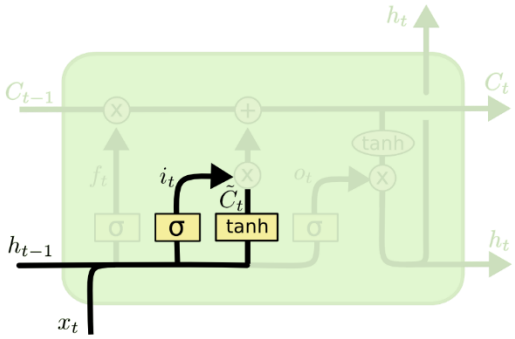
Η πρώτη πύλη είναι η πύλη επιλεκτικής συγκράτησης (forget gate). Πρόκειται για σιγμοειδές επίπεδο όπως φαίνεται στην Εικόνα 9, που αποφασίζει εάν η τιμή εξόδου της προηγούμενης κατάστασης θα κρατηθεί ή όχι.



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Εικόνα 9 Πύλη επιλεκτικής συγκράτησης (Forget Gate)

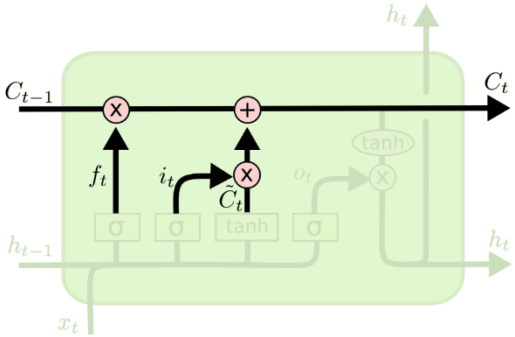
Το επόμενο βήμα είναι να γίνει η απόφαση σχετικά με το ποια νέα πληροφορία πρόκειται να καταγραφεί στο κελί μνήμης. Αυτό υλοποιείται μέσω της πύλης εισόδου (Input Gate) μεσω ενός σιγμοειδές επιπέδου και ενός εφαπτομενικού επιπέδου που παράγει υποψήφιες τιμές που θα μπορούσαν να εισαχθούν στην κατάσταση (Εικόνα 10). Ο συνδυασμός τους πρόκειται να ενημερώσει την κατάσταση του κελιού μνήμης.



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

Εικόνα 10 Πύλη Εισόδου (Input Gate)

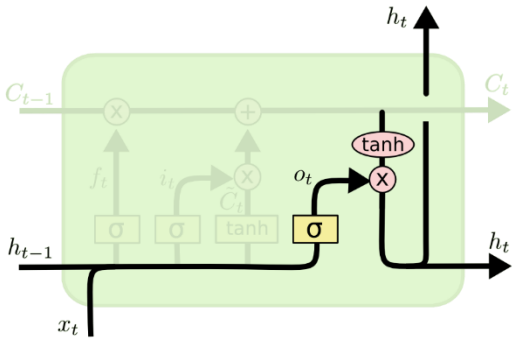
Στο σημείο αυτό γίνεται ενημέρωση της προηγούμενης κατάστασης κελιού στην νέα κατάσταση. Πολλαπλασιάζεται η προηγούμενη κατάσταση με την τιμή που προέκυψε από την πύλη επιλεκτικής συγκράτησης (Εικόνα 11) .



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Εικόνα 11 Ενημέρωση κατάστασης κελιού (Cell State)

Τελικά το αποτέλεσμα της πύλης εξόδου (Output Gate) όπως φαίνεται στην Εικόνα 12 βασίζεται στην κατάσταση του κελιού όπως προκύπτει από ένα σιγμοειδές επίπεδο και από την εφαρμογή εφαπτομενικής εξίσωσης ώστε να είναι οι τιμές στο εύρος [-1,1].



$$o_t = \sigma\left(W_o \; [\, h_{t-1}, x_t] \; + \; b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

Εικόνα 12 Πύλη Εξόδου (Output Gate)

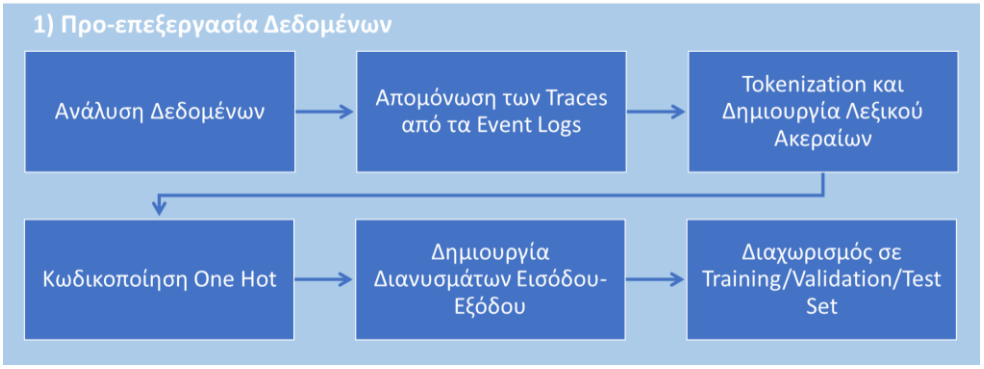
4

Η Προτεινόμενη Προσέγγιση

Η ενότητα αυτή εισαγάγει και εξηγεί βήμα-βήμα τη μεθοδολογία που ακολούθησα για την πρόβλεψη των επόμενων δραστηριοτήτων στις επιχειρηματικές διεργασίες, χρησιμοποιώντας δεδομένα από event logs, τα οποία προήλθαν από την εκτέλεση των διεργασιών αυτών. Η μεθοδολογία αυτή είναι βασισμένη στη δημιουργία ενός LSTM νευρωνικού δικτύου και αποτελείται, όπως φαίνεται και στην Εικόνα 13, από 3 φάσεις: 1) Προ-επεξεργασία Δεδομένων, 2) Υλοποίηση LSTM μοντέλου και 3) Εξαγωγή Αποτελεσμάτων.

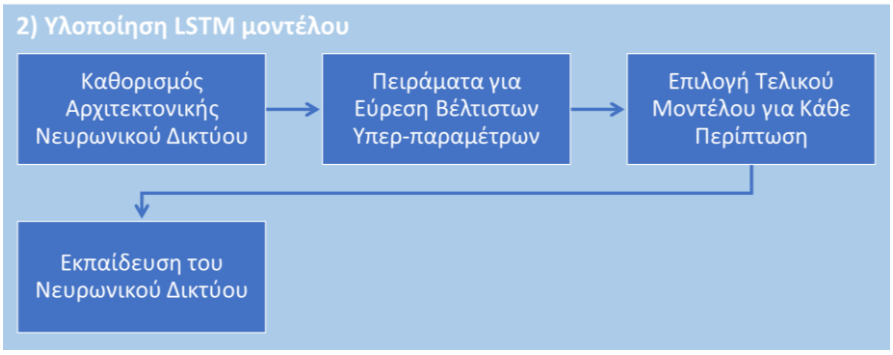
Εικόνα 13 Οι 3 φάσεις της μεθοδολογίας πρόβλεψης των επόμενων δραστηριοτήτων στις επιχειρηματικές διεργασίες

Η πρώτη φάση της μεθοδολογίας, δηλαδή η προ-επεξεργασία των event logs αποτελείται από τα στάδια που φαίνονται στην Εικόνα 14. Κατά τη φάση αυτή, τις οποίας τα στάδια θα αναλύσω με περαιτέρω λεπτομέρειες στην συνέχεια, ουσιαστικά προετοιμάζω τα δεδομένα και τα μετασχηματίζω στην κατάλληλη μορφή έτσι ώστε να εισαχθούν σωστά στο LSTM μοντέλο.



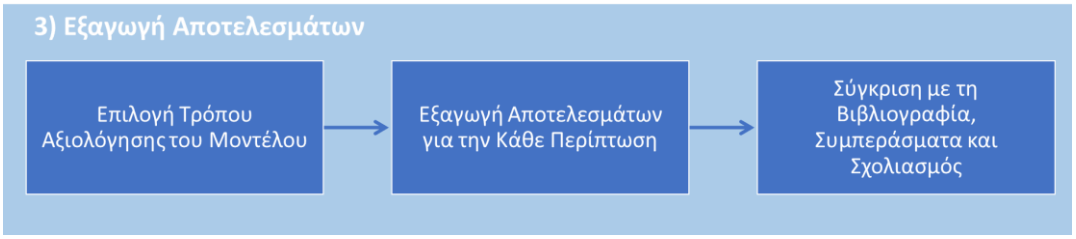
Εικόνα 14 Η πρώτη φάση της μεθοδολογίας: Προ-επεξεργασία δεδομένων

Η δεύτερη φάση της μεθοδολογίας, η οποία χωρίστηκε στα στάδια που στην Εικόνα 15 είναι η κατασκευή της αρχιτεκτονικής του μοντέλου LSTM το οποίο θα προβλέψει τις επόμενες δραστηριότητες των event log που έχουμε. Στο κάθε στάδιο ξεχωριστά θα αναφερθώ με λεπτομέρειες στις επόμενες ενότητες.



Εικόνα 15 Η δεύτερη φάση της μεθοδολογίας: Υλοποίηση LSTM μοντέλου

Η τρίτη και τελευταία φάση της μεθοδολογίας είναι η αξιολόγηση του μοντέλου και η εξαγωγή των αποτελεσμάτων. Αφού γίνει η εξαγωγή των αποτελεσμάτων του μοντέλου θα γίνει και η σύγκριση με άλλα μοντέλα της βιβλιογραφίας τα οποία αναφέρθηκαν στην ενότητα 2 και θα αναγνωριστούν οι βελτιώσεις ή οι καινούριες ιδέες για μελλοντική δουλειά. Τα στάδια που ακολούθησα για αυτήν τη φάση φαίνονται στην Εικόνα 16.



Εικόνα 16 Η τρίτη φάση της μεθοδολογίας: Εξαγωγή Αποτελεσμάτων

4.1 Προ-επεξεργασία Event Log

4.1.1 Απομόνωση των ιχνών των event logs

Αρχικά, με σκοπό τη διαπίστωση των καταλληλότερων χαρακτηριστικών (attributes) που θα εξαχθούν από το event log γίνεται λεπτομερής ανάλυση των δεδομένων. Στα πλαίσια της διπλωματικής χρησιμοποιήθηκαν για τις προβλέψεις οι δραστηριότητες (activities) του event log. Το επόμενο βήμα της προ-επεξεργασίας είναι να απομονώσουμε τα traces με τα αντίστοιχα τους events. Με τη βοήθεια ενός python script προτείνω να δημιουργηθεί ένα αρχείο κειμένου (txt) που σε κάθε γραμμή θα έχει τις δραστηριότητες ενός trace διατηρώντας την σειρά

εμφάνισης τους. Στο τέλος της κάθε γραμμής τοποθετείτε και η λέξη END για να σηματοδοτεί το που τελειώνει το κάθε trace.

Για παράδειγμα στο event log που έδωσα ως παράδειγμα στο δεύτερο κεφάλαιο η προ-επεξεργασία αυτή θα εφαρμοζόταν με τον τρόπο που φαίνεται πιο κάτω (Εικόνα 17):

Case Id	Activity
5781	X-ray
5782	Blood test
5781	Blood test
5781	CT scan
5781	Surgery
5782	Payment
5783	Check in
5783	Blood Test

X-ray, Blood Test, CT scan, Surgery, END

Blood test, Payment, END

Check in, Blood Test, END

Εικόνα 17 Παράδειγμα απομόνωσης των ιχνών από ένα event log

4.1.2 Διαχείριση Κατηγορικών Δεδομένων

Δεδομένου ότι οι υπολογιστές δεν μπορούν να επεξεργαστούν κατηγορικά δεδομένα καθώς αυτές οι κατηγορίες στις οποίες χωρίζονται δεν έχουν κάποιο νόημα γι' αυτούς, οι πληροφορίες αυτές πρέπει να προετοιμαστούν διαφορετικά εάν θέλουμε ένας υπολογιστής να μπορεί να τις επεξεργαστεί. Οπότε ένα μεγάλο κομμάτι της προ-επεξεργασίας των δεδομένων είναι η κωδικοποίηση, η οποία αντιστοιχεί κάθε κομμάτι δεδομένων με τρόπο που ο υπολογιστής μπορεί να κατανοήσει (η λέξη σημαίνει κυριολεκτικά "μετατροπή σε κώδικα υπολογιστή").

Τα μοντέλα μηχανικής μάθησης απαιτούν όλες οι μεταβλητές εισόδου και εξόδου να είναι αριθμητικές. Αυτό σημαίνει ότι εάν τα δεδομένα περιέχουν κατηγορικά δεδομένα, πρέπει να μετατραπούν σε αριθμούς αφού σκοπεύουμε να τα εισάγουμε ως είσοδο σε LSTM αλγόριθμο βαθιάς μηχανικής μάθησης.

Μια κατηγορική μεταβλητή είναι μια μεταβλητή της οποίας οι τιμές λαμβάνουν την τιμή κάποιων ετικετών (labels). Για παράδειγμα, η μεταβλητή μπορεί να είναι "χρώμα" και μπορεί να έχει τις τιμές "κόκκινο", "πράσινο" και "μπλε".

4.1.2.1 Κωδικοποίηση Ordinal- Tokenization

Μερικές φορές, τα κατηγορικά δεδομένα ενδέχεται να έχουν μια ταξινομημένη σχέση μεταξύ των κατηγοριών, όπως "πρώτο", "δεύτερο" και "τρίτο". Αυτός ο τύπος κατηγορικών δεδομένων αναφέρεται ως ordinal και οι πρόσθετες πληροφορίες της φυσικής σειράς των δεδομένων μπορεί να είναι χρήσιμες.

Πρώτο βήμα της κωδικοποίησης αυτής, προτείνω να είναι το tokenization των διαδοχικών δραστηριοτήτων στο αρχείο κειμένου με τη χρήση της κλάσης tf.keras.preprocessing.text.Tokenizer της python. Αυτή η κλάση επιτρέπει τη διανυσματοποίηση ενός σώματος κειμένου, μετατρέποντας κάθε κείμενο σε μια ακολουθία ακεραίων (κάθε ακέραιος είναι ο δείκτης ενός token στο λεξικό). Έτσι κάθε δραστηριότητα που υπάρχει στα ίχνη (traces) των δεδομένων μας θα αντιστοιχηθεί με ένα ακέραιο αριθμό και θα τοποθετηθεί με τη σειρά που εμφανίζεται σε μία λίστα.

Στην πραγματικότητα όμως, χρησιμοποιώντας αυτή την κωδικοποίηση επιτρέπουμε στο μοντέλο να υιοθετήσει μια φυσική σειρά μεταξύ των κατηγοριών και να τείνει να δίνει στους υψηλότερους αριθμούς υψηλότερα βάρη. Αυτό μπορεί να οδηγήσει σε κακή απόδοση ή απροσδόκητα αποτελέσματα από τη στιγμή που η φυσική αύξουσα σειρά δεν χαρακτηρίζει τα δεδομένα μας. Ένα παράδειγμα της κωδικοποίησης ordinal φαίνεται στην Εικόνα 18.

Χρώμα	**Κωδικοποίηση Ordinal**
Κόκκινο	1
Μπλε	2
Κίτρινο	3
Μπλε	2
Κίτρινο	3

Εικόνα 18 Παράδειγμα ordinal κωδικοποίησης

4.1.2.2 Κωδικοποίηση One Hot

Με σκοποί να μην επιτρέψουμε στο μοντέλο να υιοθετήσει μια φυσική σειρά μεταξύ των κατηγοριών και να τείνει να δίνει στους υψηλότερους αριθμούς υψηλότερα βάρη, μια άλλη κωδικοποίηση που ονομάζεται one-hot encoding μπορεί να εφαρμοστεί στην ακέραια αναπαράσταση. Η κωδικοποίηση One Hot είναι κατάλληλη για κατηγορικά δεδομένα όπου δεν υπάρχει σχέση μεταξύ των κατηγοριών.

Εδώ αφαιρείται η κωδικοποιημένη ακέραια μεταβλητή και προστίθεται μια νέα δυαδική μεταβλητή για κάθε μοναδική ακέραια τιμή. Περιλαμβάνει ένα διανυσματικό τύπο αναπαράστασης στον οποίο όλα τα στοιχεία ενός διανύσματος είναι 0, εκτός από ένα, το οποίο έχει τιμή 1, και το 1 αντιπροσωπεύει μια από τις κατηγορίες των δεδομένων. Έτσι, κωδικοποιούμε την λίστα με τους ακέραιους σε ένα πίνακα με τις one hot αναπαραστάσεις. Επίσης, εάν η κατηγορική μεταβλητή είναι μεταβλητή εξόδου, θα μπορούσαμε να μετατρέψουμε τις τιμές πίσω στην κατηγορική τους μορφή για να τις παρουσιάσουμε στην εφαρμογή μας.

Για παράδειγμα, αν η μεταβλητή μας ήταν "χρώμα" και οι ετικέτες ήταν "κόκκινο", "κίτρινο" και "πράσινο", θα κωδικοποιούμε καθεμία από αυτές τις ετικέτες ως δυαδικό διάνυσμα τριών στοιχείων ως εξής (Εικόνα 19):

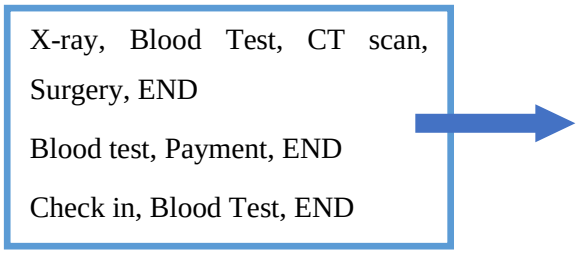
Χρώμα	**Κόκκινο**	**Μπλε**	**Κίτρινο**
Κόκκινο	1	0	0
Μπλε	0	1	0
Κίτρινο	0	0	1
Μπλε	0	1	0
Κίτρινο	0	0	1

Εικόνα 19 Παράδειγμα One Hot κωδικοποίησης

4.1.3 Διαχωρισμός σε διανύσματα εισόδου-εξόδου

Ουσιαστικά το μοντέλο LSTM που υλοποιούμε ΄θα μάθει' μια συνάρτηση με την οποία αντιστοιχεί μια ακολουθία προηγούμενων δραστηριοτήτων (είσοδος) σε μια ή περισσότερες διαδοχικές δραστηριότητες εξόδου. Ως εκ τούτου, η ακολουθία των δραστηριοτήτων πρέπει να μετατραπεί σε πολλαπλά παραδείγματα από τα οποία μπορεί να μάθει το LSTM. Μπορούμε να χωρίσουμε την ακολουθία εισόδου (ουσιαστικά το training set) σε πολλαπλά μοτίβα εισόδου/εξόδου που ονομάζονται δείγματα (samples), και όπου ένα ή περισσότερα διαδοχικά βήματα χρησιμοποιούνται ως είσοδος και 1 ή περισσότερα βήματα χρησιμοποιούνται ως έξοδος για την πρόβλεψη για την οποία εκπαιδεύεται το LSTM. Για τον σκοπό αυτό γράφτηκε ειδική συνάρτηση στην python.

Επιστρέφοντας στο προηγούμενο παράδειγμα ο διαχωρισμός των samples σε είσοδο και έξοδο εάν θέλαμε να έχουμε δύο δραστηριότητες ως είσοδο και μία ως έξοδο θα γινόταν με αυτό τον τρόπο (Εικόνα 20):



Είσοδος	Έξοδος
X-ray, Blood Test	CT scan
Blood Test, CT scan	Surgery
CT scan, Surgery	END

Εικόνα 20 Παράδειγμα διαχωρισμού σε διανύσματα εισόδου-εξόδου

4.1.4 Διαχωρισμός σε Training/Validation/Test Set

Επόμενο και αναγκαίο βήμα αποτελεί ο διαχωρισμός του συνόλου των αρχικών δεδομένων σε δύο υποσύνολα: το υποσύνολο εκπαίδευσης (Training Set) και το υποσύνολο ελέγχου (Test Set). Όπως αναφέρουν και τα ονόματα, το πρώτο αφορά την εκπαίδευση του νευρωνικού και το δεύτερο την εξέταση και αξιολόγηση των αποτελεσμάτων.

Ο διαχωρισμός αυτός πρέπει να τηρεί δύο πολύ βασικές προϋποθέσεις. Πρώτον, το υποσύνολο ελέγχου πρέπει να είναι τόσο μεγάλο ώστε να παρέχει στατιστικά σωστά αποτελέσματα ενώ ταυτόχρονα το υποσύνολο εκπαίδευσης πρέπει να είναι αρκετά μεγάλο ώστε να εκπαιδευτεί βέλτιστα το νευρωνικό αποφεύγοντας φυσικά το Overfitting. Δεύτερον πρέπει το υποσύνολο ελέγχου να είναι αντιπροσωπευτικό του συνόλου δεδομένων, με άλλα λόγια, να μην αποτελεί ένα υποσύνολο ελέγχου με διαφορετικά χαρακτηριστικά από το υποσύνολο εκπαίδευσης. Με γνώμονα τα παραπάνω για την εκπαίδευση του μοντέλου χρησιμοποιήθηκε το 67% των δεδομένων, ενώ για την αξιολόγησή του χρησιμοποιήθηκε το υπόλοιπο 33% (Εικόνα 22).

Κατά τη διάρκεια της εύρεσης των βέλτιστων υπερ παραμέτρων για την εκπαίδευση του νευρωνικού τα δεδομένα του υποσύνολο ελέγχου χωρίστηκαν σε 60% δεδομένα εκπαίδευσης (Training Set) 20% δεδομένα επικύρωσης (Validation Set) και 20% δεδομένα ελέγχου (Test Set) όπως φαίνεται στην Εικόνα 21. Το σύνολο επικύρωσης είναι υποσύνολο με γνωστές τις κλάσεις, διαφορετικό από το σύνολο εκπαίδευσης, που χρησιμοποιείται για τη ρύθμιση των παραμέτρων του αλγορίθμου. Είναι διαφορετικό σύνολο από τα δεδομένα ελέγχου τα οποία είναι υποσύνολο δειγμάτων με άγνωστες τις κλάσεις, και τα οποία χρησιμοποιούνται για την αξιολόγηση της απόδοσης του αλγόριθμου.

Εικόνα 21 Διαχωρισμός Δεδομένων στο στάδιο εύρεσης βέλτιστων υπερ-παραμέτρων

Εικόνα 22 Διαχωρισμός Δεδομένων στο στάδιο των προβλέψεων

4.2 Αρχιτεκτονική του μοντέλου

Την προ-επεξεργασία δεδομένων ακολουθεί η τελική υλοποίηση και ο καθορισμός της αρχιτεκτονικής του LSTM νευρωνικού δικτύου. Καθοριστικό ρόλο σε αυτό παίζουν οι σωστές επιλογές των υπερ παραμέτρων του μοντέλου. Η κατάλληλη επιλογή κόμβων (nodes) και επιπέδων (layers) δεν έχει συγκεκριμένο αλγόριθμο και βασίζεται αποκλειστικά στο πρόβλημα που καλούμαστε να επιλύσουμε. Το (χρονικά) διαδοχικό API (sequential) επιτρέπει τη δημιουργία μοντέλων επίπεδο προς επίπεδο. Είναι περιορισμένο, καθώς δεν επιτρέπει την δημιουργία μοντέλων που μοιράζονται επίπεδα ή έχουν πολλές εισόδους ή εξόδους, είναι όμως κατάλληλο για την φύση του δικού μας προβλήματος.

Χρειάστηκε λοιπόν να ληφθούν αποφάσεις σχετικά με την επιλογή παραμέτρων όπως ο κατάλληλος αριθμός επιπέδων (layers), ο κατάλληλος αριθμός νευρώνων (neurons), το ποσοστό του περιορισμού ενεργοποίησης (dropout), ο αλγόριθμος βελτιστοποίησης (optimizer), η συνάρτηση ενεργοποίησης (activation function) και η απώλεια (loss).

Στα πλαίσια της διπλωματικής μου προτείνω και χρησιμοποίησα το Talos για βελτιστοποίηση των υπερ-παραμέτρων του νευρωνικού μου δικτύου. Κατά την εκτέλεση του κώδικα με το Talos στην εντολή σάρωσης, όλοι οι πιθανοί συνδυασμοί υπερ-παραμέτρων δοκιμάζονται σε ένα πείραμα. Στη συνέχεια, το καλύτερο μοντέλο αποθηκεύεται και μπορεί να εφαρμοστεί ακριβώς όπως θα κάναμε και μηχανικά σε ένα νευρωνικό δίκτυο χρησιμοποιώντας το Keras. Με το Talos όπως φαίνεται και στα πειράματα της επόμενης ενότητας επέλεξα τις εξής υπερ-παραμέτρους: τον κατάλληλος αριθμό νευρώνων (neurons), το ποσοστό του περιορισμού ενεργοποίησης (dropout) και το batch size. Οι υπόλοιπες επιλέχθηκαν με τη λογική που περιγράφεται στις επόμενες παραγράφους.

Η συνάρτηση ενεργοποίησης (activation function) που θα επιλεγεί εξαρτάται με την εφαρμογή του αλγορίθμου. Επειδή το πρόβλημά μας είναι πρόβλημα ταξινόμησης (classification) και συγκεκριμένα υπάρχουν περισσότερες από δύο αμοιβαία αποκλειστικές κατηγορίες (multiclass classification), τότε το επίπεδο εξόδου θα έχει έναν κόμβο ανά κατηγορία και θα πρέπει να χρησιμοποιηθεί η συνάρτηση ενεργοποίησης softmax. Η συνάρτηση softmax εξάγει ένα διάνυσμα τιμών που συνολικά αθροίζονται στο 1,0 και μπορούν να ερμηνευτούν ως πιθανότητες συμμετοχής σε κλάση. Η είσοδος στη συνάρτηση είναι ένας φορέας πραγματικών τιμών και η έξοδος είναι ένας φορέας του ίδιου μήκους με τιμές που αθροίζονται στο 1,0 σαν πιθανότητες. Οι τιμές εισόδου μπορεί να είναι θετικές, αρνητικές, μηδενικές ή μεγαλύτερες από μία, αλλά το softmax τις μετατρέπει σε τιμές μεταξύ 0 και 1, έτσι ώστε να μπορούν να ερμηνευτούν ως πιθανότητες. Εάν μία από τις εισόδους είναι μικρή ή αρνητική, h softmax το

μετατρέπει σε μικρή πιθανότητα και εάν μια είσοδος είναι μεγάλη, τότε τη μετατρέπει σε μεγάλη πιθανότητα, αλλά θα παραμένει πάντα μεταξύ 0 και 1.

Ο τύπος της συνάρτησης softmax είναι ως εξής:

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (9)$$

όπου όλες οι τιμές z_i είναι τα στοιχεία του διανύσματος εισόδου και μπορούν να πάρουν οποιαδήποτε πραγματική τιμή. Ο όρος στο κάτω μέρος του τύπου είναι ο όρος κανονικοποίησης που διασφαλίζει ότι όλες οι τιμές εξόδου της συνάρτησης θα αθροιστούν στο 1, αποτελώντας έτσι μια έγκυρη κατανομή πιθανότητας. Η επεξήγηση της εξίσωσης της συνάρτησης softmax παρουσιάζεται στον επόμενο πίνακα (Πίνακας 3):

Πίνακας 3 Η επεξήγηση της εξίσωσης της συνάρτησης softmax

$\vec{z}$	Το διάνυσμα εισόδου στη συνάρτηση softmax που αποτελείται από (z0, ... zK)
z_i	Όλες οι τιμές zi είναι τα στοιχεία του διανύσματος εισόδου στη συνάρτηση softmax και μπορούν να πάρουν οποιαδήποτε πραγματική τιμή, θετική, μηδενική ή αρνητική. Για παράδειγμα, ένα νευρωνικό δίκτυο θα μπορούσε να έχει έξοδο ενός διανύσματος όπως (-0.62, 8.12, 2.53), ο οποίος δεν είναι έγκυρη κατανομή πιθανότητας, εξ ου και γιατί η softmax θα ήταν απαραίτητη.
e^{z_i}	Η τυπική εκθετική συνάρτηση εφαρμόζεται σε κάθε στοιχείο του διανύσματος εισόδου. Αυτό δίνει θετική τιμή πάνω από 0, η οποία θα είναι πολύ μικρή αν η είσοδος ήταν αρνητική και πολύ μεγάλη αν η είσοδος ήταν μεγάλη. Ωστόσο, εξακολουθεί να μην είναι σταθερό στο εύρος (0, 1) που απαιτείται για να αποτελεί πιθανότητα.
$\sum_{j=1}^{K} e^{z_j}$	Ο όρος στο κάτω μέρος του τύπου είναι ο όρος κανονικοποίησης. Διασφαλίζει ότι όλες οι τιμές εξόδου της συνάρτησης θα αθροιστούν στο 1 και κάθε μία θα βρίσκεται στο εύρος (0, 1), αποτελώντας έτσι μια έγκυρη κατανομή πιθανότητας.
K	Ο αριθμός κλάσεων στον ταξινομητή πολλαπλών κλάσεων (multi-class classification).

Η συνάρτηση απωλειών (loss function) και η συνάρτηση ενεργοποίησης (activation function) επιλέγονται συχνά μαζί. Η συνάρτηση απωλειών Cross Entropy χρησιμοποιείται πολύ συχνά για προβλήματα ταξινόμησης πολλαπλών κατηγοριών. Προορίζεται για χρήση με ταξινόμηση πολλαπλών κατηγοριών, όπου οι τιμές-στόχοι είναι στο σύνολο {0, 1, 3,…, n}, όπου σε κάθε κατηγορία έχει μια μοναδική ακέραια τιμή. Μαθηματικά, είναι η προτιμώμενη συνάρτηση

απώλειας στο πλαίσιο συμπερασμάτων μέγιστης πιθανότητας. Είναι η συνάρτηση απώλειας που πρέπει να αξιολογηθεί πρώτα και να αλλάξει μόνο υπάρχει πολύ καλός λόγος.

Ονομάζεται επίσης λογαριθμική απώλεια, απώλεια καταγραφής ή λογιστική απώλεια. Κάθε προβλεπόμενη πιθανότητα κλάσης συγκρίνεται με την πραγματική επιθυμητή έξοδο τάξης 0 ή 1 και υπολογίζεται μια βαθμολογία / απώλεια που τιμωρεί την πιθανότητα βάσει του πόσο μακριά είναι από την πραγματική αναμενόμενη τιμή. Το πέναλτι είναι λογαριθμικό στη φύση αποδίδοντας μεγάλο σκορ για μεγάλες διαφορές κοντά στο 1 και μικρό σκορ για μικρές διαφορές που τείνουν στο 0. Η απώλεια εγκάρσιας εντροπίας χρησιμοποιείται κατά την προσαρμογή των βαρών του μοντέλου κατά τη διάρκεια της εκπαίδευσης. Ο στόχος είναι να ελαχιστοποιηθεί η απώλεια, δηλαδή όσο μικρότερη είναι η απώλεια τόσο καλύτερο είναι το μοντέλο. Ένα τέλειο μοντέλο έχει απώλεια εγκάρσιας εντροπίας 0.

Η συνάρτηση απώλειας Cross-Entropy ορίζεται ως:

$$L_{CE} = -\sum_{i=1}^{n} t_i \log(p_i) \ \ για\ n\ κλάσεις \tag{10}$$

Όπου t_i η τιμή στόχος και p_i είναι η πιθανότητα της συνάρτησης softmax για την i κλάση.

Η Cross Entropy μπορεί να οριστεί ως η συνάρτηση απώλειας στο Keras, καθορίζοντας την «categorical_crossentropy» κατά τη σύνταξη του μοντέλου.

Για την επιλογή του βελτιστοποιητή (optimizer), ο adaptive moment estimation (adam) έχει αποδειχθεί ότι λειτουργεί καλά στις περισσότερες πρακτικές εφαρμογές και λειτουργεί καλά με λίγες μόνο αλλαγές στις υπερ-παραμέτρους. Ο αλγόριθμος αυτός αποτελεί μια προέκταση του αλγορίθμου SGD (Stochastic Gradient Descent) και έχει υιοθετηθεί εκτενώς τα τελευταία χρόνια για εφαρμογές Deep Learning. Σύμφωνα με τους δημιουργούς του, μερικά από τα πολυάριθμα πλεονεκτήματά του είναι τα εξής: άμεση και εύκολη υλοποίηση, υπολογιστική αποδοτικότητα, μικρές απαιτήσεις μνήμης, υψηλή απόδοση για προβλήματα που χαρακτηρίζονται από πολλά δεδομένα και παραμέτρους, ελάχιστες απαιτήσεις στην προσαρμογή των υπερπαραμέτρων του δικτύου.

Ο τρόπος με τον οποίο λειτουργεί ο adam είναι διαφορετικός σε σχέση με τον SGD όσον αφορά το ρυθμό εκπαίδευσης. Στον αλγόριθμο SGD διατηρείται ένας σταθερός ρυθμός εκπαίδευσης για την ενημέρωση όλων των βαρών και για όλη τη φάση της εκπαίδευσης. Στον

αλγόριθμο Adam κρατείται ένας ρυθμός εκπαίδευσης για κάθε παράμετρο (βάρος) του δικτύου και υιοθετείται χωριστά στα βάρη όσο προχωράει η εκπαίδευση. Με άλλα λόγια υπολογίζει μεμονωμένους ρυθμούς εκπαίδευσης για κάθε παράμετρο και το κάνει προσεγγίζοντας τις ροπές πρώτης και δεύτερης τάξης των παραγώγων (gradients). Στη θεωρία των πιθανοτήτων

και της στατιστικής η ροπή πρώτης τάξης είναι η μέση τιμή και η ροπή δεύτερης τάξης είναι η διασπορά.

Επόμενο κομμάτι της υλοποίησης αποτελεί η επιλογή του πλήθους των εποχών (Epochs) και του Batch Size. Οι Epochs είναι μια υπερ-παράμετρος που καθορίζει τον αριθμό των φορών που ο αλγόριθμος εκμάθησης θα λειτουργεί σε ολόκληρο το σύνολο δεδομένων εκπαίδευσης. Το Batch size είναι μια υπερ-παράμετρος που καθορίζει τον αριθμό των δειγμάτων που πρέπει να επεξεργαστούν από το σύστημα πριν από την ενημέρωση των εσωτερικών παραμέτρων του μοντέλου. Η τελική επιλογή των περαμέτρων του νευρωνικού δικτύου παρουσιάζεται στον πίνακα που ακολουθεί (Πίνακας 4).

Πίνακας 4 Επιλογή παραμέτρων νευρωνικού δικτύου

Συνάρτηση ενεργοποίησης (activation function)	Softmax
Συνάρτηση απωλειών (loss function)	Categorical Cross-entropy
Βελτιστοποιητής (optimizer)	adam

4.3 Μέθοδος Αξιολόγησης του μοντέλου

Η επιλογή της σωστής μετρικής είναι ζωτικής σημασίας κατά την αξιολόγηση των μοντέλων μηχανικής μάθησης. Η ταξινόμηση (classification) είναι ένα από τα πιο διαδεδομένα προβλήματα στη μηχανική μάθηση με διάφορες βιομηχανικές εφαρμογές, από αναγνώριση προσώπου, κατηγοριοποίηση βίντεο, εποπτεία περιεχομένου, ιατρική διάγνωση, έως ταξινόμηση κειμένου, ανίχνευση ρητορικής μίσους στο Twitter. Το δικό μου θέμα στη διπλωματική αποτελεί επίσης ένα πρόβλημα ταξινόμησης. Υπάρχουν διάφοροι τρόποι για την αξιολόγηση ενός μοντέλου ταξινόμησης και το Keras προσφέρει πολλαπλές επιλογές για μέτρηση της ακρίβειας.

Όλες οι μετρήσεις αξιολόγησης για ένα μοντέλο ταξινόμησης πολλαπλών κλάσεων μπορούν να γίνουν κατανοητές στο πλαίσιο ενός μοντέλου δυαδικής ταξινόμησης (όπου οι τάξεις είναι απλώς «θετικές» και «αρνητικές»). Αυτές οι μετρήσεις προέρχονται από τις ακόλουθες τέσσερις κατηγορίες:

1. True Positives (TP): Στοιχεία όπου η πραγματική ετικέτα είναι θετική και των οποίων η κλάση προβλέπεται σωστά να είναι θετική.
2. False Positives (FP): Στοιχεία όπου η πραγματική ετικέτα είναι αρνητική και των οποίων η τάξη έχει προβλεφθεί εσφαλμένα ότι είναι θετική.
3. True Negatives (TN): Στοιχεία όπου η πραγματική ετικέτα είναι αρνητική και των οποίων η κλάση προβλέπεται σωστά ως αρνητική.

4. False Negatives (FN): Στοιχεία όπου η πραγματική ετικέτα είναι θετική και των οποίων η τάξη έχει προβλεφθεί εσφαλμένα ότι είναι αρνητική.

Μπορούμε να κατανοήσουμε ένα πρόβλημα ταξινόμησης πολλών κλάσεων ως ένα σύνολο πολλών δυαδικών προβλημάτων ταξινόμησης - ένα για κάθε τάξη. Για παράδειγμα, στην περίπτωση του παραδείγματος με τα χρώματα που αναφέρθηκε πιο πάνω, όταν εξετάζουμε την κατηγορία "Κόκκινο", ένα αληθινό θετικό (true positive) συμβαίνει όταν το χρώμα που στην πραγματικότητα είναι κόκκινο προβλέπεται να είναι κόκκινο. Οποιαδήποτε άλλη πρόβλεψη (μπλε ή κίτρινο) θα θεωρείται ψευδώς αρνητική (false negative). Αυτό ισχύει για κάθε κατηγορία: αυτό που αποκαλούμε «θετικό» και «αρνητικό» θα αλλάξει ανάλογα με την πραγματική ετικέτα του αντικειμένου. Αυτό σημαίνει ότι υπάρχουν πολλές κατηγορίες που θεωρούνται αληθινά αρνητικά για μια δεδομένη πρόβλεψη

Ο πίνακας Confusion (confusion matrix) είναι ένας πίνακας N x N που χρησιμοποιείται για την αξιολόγηση της απόδοσης ενός μοντέλου ταξινόμησης, όπου N είναι ο αριθμός των κατηγοριών στόχων. Στον πίνακα συγκρίνουμε τις πραγματικές τιμές-στόχους με εκείνες που προβλέπονται από το μοντέλο μηχανικής μάθησης. Αυτό μας δίνει μια ολιστική εικόνα για το πόσο καλά αποδίδει το μοντέλο ταξινόμησης και τι είδους λάθη κάνει.

Στον άξονα x έχουμε τις προβλεπόμενες ετικέτες και στον άξονα y έχουμε τις πραγματικές ετικέτες των δειγμάτων . Στην ιδανική περίπτωση, ένας τέλειος ταξινομητής θα οδηγούσε σε έναν πίνακα confusion όπου έχουμε τιμές μόνο στη διαγώνιο, δηλαδή σε αυτή την περίπτωση ταξινομούμε σωστά όλα τα δείγματα και για τις 3 κλάσεις / ομάδες που έχουμε.

Επιστρέφοντας στο παράδειγμα με τα χρώματα , ο προβλέπτης μας πρέπει να προβλέψει ποιο χρώμα εμφανίζεται σε κάθε φωτογραφία. Αυτό είναι ένα πρόβλημα ταξινόμησης με κλάσεις N= 3. Κοιτάζοντας τη δεύτερη σειρά στην Εικόνα 23 που αναφέρεται στην ομάδα μπλε, μπορούμε να δούμε ότι ταξινομήσαμε σωστά 2 δείγματα (από το σύνολο των 10) και χάσαμε 1 δείγμα μπλε που είχε προβλεφθεί εσφαλμένα ως «κόκκινο».

		Πραγματική τιμή			
		Κόκκινο	Μπλε	Κίτρινο	
Πρόβλεψη	Κόκκινο	4	6	3	Συνολικά προβλεπόμενα κόκκινα=13
	Μπλε	1	2	0	Συνολικά προβλεπόμενα μπλε=3
	Κίτρινο	1	2	6	Συνολικά προβλεπόμενα κίτρινα=9
		Συνολικά Αληθινά κόκκινα=6	Συνολικά Αληθινά μπλε=10	Συνολικά Αληθινά κίτρινα=9	**Σύνολο=25**

Εικόνα 23 Πάραδειγμα πίνακα Confusion

Σε πολλές περιπτώσεις, η αξιολόγηση της απόδοσης (accuracy) των μοντέλων από μια πιο σφαιρική οπτική θα είναι η καλύτερη επιλογή ερμηνείας, καθώς και επαρκής ως προς την απόδοση του μοντέλου. Η ακρίβεια ταξινόμησης είναι ίσως οι απλούστερη μετρική που μπορεί κανείς να φανταστεί και ορίζεται ως ο αριθμός των σωστών προβλέψεων διαιρούμενος με τον συνολικό αριθμό προβλέψεων, πολλαπλασιασμένος επί 100.

Η ακρίβεια σε ένα μοντέλο ταξινόμησης πολλαπλών κλάσεων μπορεί να οριστεί ως:

$$Accuracy = \frac{\sum_{i=1}^{i=N} TP_i}{\sum_{i=1}^{i=N} TP_i + FP_i} = \frac{Αριθμός\ σωστών\ προβλέψεων}{Αριθμός\ συνολικών\ προβλέψεων} \quad (11)$$

Ακόμα μία χρήσιμη μετρική για την αξιολόγηση του μοντέλου είναι το Precision. Το Precision εκφράζει το ποσοστό των προβλέψεων που το μοντέλο μας λέει ότι είναι θετικά και είναι πραγματικά θετικά. Με άλλα λόγια, το Precision μας λέει πόσο μπορούμε να εμπιστευτούμε το μοντέλο όταν προβλέπει κάτι ως θετικό.

Μπορεί να υπολογιστεί για κάθε κλάση k ως:

$$Precision = \frac{TP_k}{TP_k + FP_k} \quad (12)$$

Για παράδειγμα, το Precision για την κατηγορία "Κόκκινο", είναι ο αριθμός των σωστών προβλεπόμενων κόκκινων (4) από τα συνολικά προβλεπόμενα κόκκινα (4 + 3 + 6 = 13), που ανέρχεται σε 4/13 = 30,8%. Έτσι λοιπόν, μόνο περίπου το 1/3 των χρωμάτων που η πρόβλεψή μας χαρακτηρίζει ως κόκκινα είναι στην πραγματικότητα κόκκινα.

Το Recall μετρά την προγνωστική ακρίβεια του μοντέλου για τη θετική κλάση: διαισθητικά, μετρά την ικανότητα του μοντέλου να εντοπίζει όλες τις θετικές μονάδες στο σύνολο δεδομένων.

Μπορεί να υπολογιστεί για κάθε κλάση k ως:

$$Recall = \frac{TP_k}{TP_k + FN_k} \tag{13}$$

Πίσω στο παράδειγμα, το recall για την κατηγορία "Κόκκινο", είναι ο αριθμός των σωστά προβλεπόμενων κόκκινων χρωμάτων (4) από τον αριθμό των πραγματικών κόκκινων χρωμάτων (4 + 1 + 1 = 6), που είναι 4/6 = 66,7%. Αυτό σημαίνει ότι ο ταξινομητής μας ταξινόμησε τα 2/3 των κόκκινων χρωμάτων ως κόκκινα.

Με τους πιο πάνω τύπους υπολογίζουμε το Precision και Recall της κάθε κλάσης ξεχωριστά.

Πώς μπορούμε λοιπόν να υπολογίσουμε το συνολικό Precision και Recall των μοντέλων ταξινόμησης πολλαπλών κατηγοριών; Προκειμένου να απαντηθεί αυτό τι ερώτημα, χρησιμοποιούμε μεθόδους macro και micro averaging.

Η μετρική Μίκρο ακρίβειας (Micro precision) και Μίκρο Ανάκλησης (Micro Recall) υπολογίζεται από τα αληθινά θετικά (TP) των ξεχωριστών κλάσεων, τα αληθινά αρνητικά (TN), τα ψευδώς θετικά (FP) και τα ψευδώς αρνητικά (FN) του μοντέλου. Η ιδέα είναι να εξετάσουμε όλες τις μονάδες μαζί, χωρίς να λαμβάνουμε υπόψη πιθανές διαφορές μεταξύ των κλάσεων. Επομένως, η ακρίβεια μικρο-μέσου όρου υπολογίζεται ως εξής:

$$MicroAveragePrecision = \frac{\sum_{k=1}^{K} TP_k}{\sum_{k=1}^{K} TP_k + \sum_{k=1}^{K} FP_k} \tag{14}$$

$$MicroAverageRecall = \frac{\sum_{k=1}^{K} TP_k}{\sum_{k=1}^{K} TP_k + \sum_{k=1}^{K} FN_k} \tag{15}$$

Οι πιο πάνω εξισώσεις είναι τελικά ίσες αφού στην ταξινόμηση πολλών κλάσεων ισχύει για τις ψευδείς περιπτώσεις: $\sum_{k=1}^{K} FN_k = \sum_{k=1}^{K} FP_k$

Η μετρική Μάκρο ακρίβειας (Macro Precision) και Μάκρο Ανάκλησης (Macro Recall) υπολογίζεται ως αριθμητικός μέσος όρος βαθμολογίας ακρίβειας και ανάκλησης των μεμονωμένων κλάσεων.

$$MacroAveragePrecision = \frac{\sum_{k=1}^{K} Precision_k}{K} \quad (16)$$

$$MacroAverageRecall = \frac{\sum_{k=1}^{K} Recall_k}{K} \quad (17)$$

Τέλος μπορούμε να υπολογίσουμε το weighted average των Precision και Recall το οποίο είναι το άθροισμα των βαθμολογιών όλων των κλάσεων μετά τον πολλαπλασιασμό των αντίστοιχων αναλογιών της κάθε κλάσης στο σύνολο των δεδομένων.

5

Υλοποίηση

5.1 Εργαλεία και Βιβλιοθήκες

Η γλώσσα προγραμματισμού που χρησιμοποιήθηκε για την κατασκευή του αλγορίθμου μας είναι η python και η πλατφόρμα στην οποία έγινε ο προγραμματισμός είναι το Jupyter Notebook. Το Jupyter Notebook είναι μια διαδικτυακή εφαρμογή ανοικτού κώδικα που επιτρέπει τη δημιουργία και τον διαμοιρασμό εγγράφων που περιέχουν ζωντανό κώδικα, εξισώσεις, απεικονίσεις και αφηγηματικό κείμενο. Η Python αποτελεί μια ερμηνευμένη, υψηλού επιπέδου, γλώσσα προγραμματισμού γενικού σκοπού. Η φιλοσοφία σχεδιασμού της Python δίνει έμφαση στην αναγνωσιμότητα του κώδικα με την αξιοσημείωτη χρήση σημαντικού κενού χώρου. Οι γλωσσικές κατασκευές και η αντικειμενοστραφής προσέγγιση στοχεύουν να βοηθήσουν τους προγραμματιστές να γράψουν σαφή, λογικό κώδικα για μικρά και μεγάλα έργα. Είναι dynamically typed και garbage-collected. Υποστηρίζει πολλαπλά παραδείγματα προγραμματισμού, συμπεριλαμβανομένων του δομημένου (ιδιαίτερα διαδικαστικού), αντικειμενοστραφούς και λειτουργικού προγραμματισμού. Φυσικά ενδείκνυται για την επίλυση προβλημάτων που κατηγοριοποιούνται ομοίως με το δικό μας πρόβλημα.

Τα πλαίσια (frameworks) διαδραματίζουν καθοριστικό ρόλο στον τομέα της επιστήμης δεδομένων. Αποτελούν μια συλλογή πακέτων και βιβλιοθηκών που βοηθούν στην απλοποίηση της συνολικής εμπειρίας προγραμματισμού για τη δημιουργία ενός συγκεκριμένου είδους εφαρμογών. Το Keras και το TensorFlow είναι από τα πιο δημοφιλή framework όσον αφορά τη βαθιά μάθηση.

Το TensorFlow είναι μια ολοκληρωμένη πλατφόρμα ανοικτού κώδικα για μηχανική μάθηση. Είναι ένα ολοκληρωμένο και ευέλικτο οικοσύστημα εργαλείων, βιβλιοθηκών και άλλων πόρων που παρέχουν ροές εργασίας με API υψηλού επιπέδου. Προσφέρει διάφορα επίπεδα εννοιών

για να επιλέξει ο χρήστης αυτό που χρειάζεται για να δημιουργήσει και να αναπτύξει μοντέλα μηχανικής μάθησης. Μερικά από τα εμφανή χαρακτηριστικά είναι η εύκολη κατασκευή μοντέλων, καθώς το TensorFlow προσφέρει πολλαπλά επίπεδα αφαίρεσης για κατασκευή και εκπαίδευση μοντέλων, στιβαρή παραγωγή ML οπουδήποτε, αφού επιτρέπει την εκπαίδευση και εύκολη ανάπτυξη του μοντέλου, ανεξάρτητα από τη γλώσσα ή την πλατφόρμα που χρησιμοποιεί ο χρήστης, και τέλος προσφέρει ισχυρό πειραματισμό για έρευνα. Το TensorFlow δίνει την ευελιξία και τον έλεγχο, καθώς και δυνατότητες όπως το Keras Functional API για τη δημιουργία σύνθετων τοπολογιών.

Το Keras, από την άλλη πλευρά, είναι μια βιβλιοθήκη νευρωνικών δικτύων υψηλού επιπέδου που λειτουργεί στην κορυφή των TensorFlow, CNTK και Theano. Η χρήση του Keras στη βαθιά μάθηση επιτρέπει την εύκολη και γρήγορη δημιουργία πρωτότυπων μοντέλων νευρωνικών δικτύων στην έρευνα καθώς και την απρόσκοπτη λειτουργία σε CPU και GPU. Τα κύρια πλεονεκτήματα του Keras είναι πως είναι φιλικό προς το χρήστη, διότι το Keras έχει μια απλή, συνεπή διεπαφή, βελτιστοποιημένη για κοινές περιπτώσεις χρήσης, η οποία παρέχει σαφή και ενεργή ανατροφοδότηση για σφάλματα χρήστη. Τα μοντέλα στο Keras κατασκευάζονται συνδέοντας διαμορφώσιμα δομικά στοιχεία μαζί, με λίγους περιορισμούς. Προσφέρει σταθερά και απλά API που βοηθούν στην ελαχιστοποίηση του αριθμού των ενεργειών χρήστη που απαιτούνται για συνήθεις περιπτώσεις χρήσης, καθώς επίσης παρέχει σαφή και ενεργή ανατροφοδότηση σχετικά με το σφάλμα χρήστη. Για όλους τους παραπάνω λόγους επιλέχθηκαν τα ανωτέρω Framework για την υλοποίηση.

Ένα ακόμα χρήσιμο εργαλείο που θα χρησιμοποιήσω για τον συντονισμό των υπερ-παραμέτρων (hyperparameter tuning) του νευρωνικού μου δικτύου είναι το Talos. Το Talos κυκλοφόρησε στις 11 Μαΐου 2018 και έκτοτε έχει αναβαθμιστεί επτά φορές. Λειτουργεί για Python 2 και Python 3, και ακολουθεί μια ροή εργασίας POD (Prepare, Optimize, Deploy), για να δημιουργήσει έναν ευέλικτο και αποτελεσματικό αγωγό με αποτελέσματα πρόβλεψης τελευταίας τεχνολογίας. Κατά την εκτέλεση του κώδικα με το Talos στην εντολή σάρωσης, όλοι οι πιθανοί συνδυασμοί υπερ-παραμέτρων δοκιμάζονται σε ένα πείραμα. Στη συνέχεια, το καλύτερο μοντέλο αποθηκεύεται και μπορεί να εφαρμοστεί ακριβώς όπως θα κάναμε και μηχανικά σε ένα νευρωνικό δίκτυο χρησιμοποιώντας το Keras.

5.2 Τα Δεδομένα

Το σύνολο δεδομένων που χρησιμοποιήθηκε στη διπλωματική πάρθηκε από το Business Process Intelligence Challenge του 2017 (BPIC'17) και περιγράφει τη διαδικασία υποβολής αιτήσεων δανείου σε ένα Ολλανδικό Οικονομικό Ινστιτούτο. Περιλαμβάνει δύο διαφορετικά

σύνολα δεδομένων που περιέχουν δεδομένα καταγραφής γεγονότων (event logs) τόσο για τη διαδικασία αίτησης (Application Event Log) όσο και για τις διαδικασίες δημιουργίας προσφορών δανείων (Offer Event Log).

Το σύνολο δεδομένων περιέχει δεδομένα καταγραφής γεγονότων από την περίοδο μεταξύ 01/01/2016 και 02/02/2017. Επιπλέον, και για τα δύο αρχεία καταγραφής γεγονότων υπάρχει ένα επιπλέον μοναδικό αναγνωριστικό ID και έτσι κάθε γεγονός μπορεί να αναγνωριστεί μοναδικά όχι μόνο στο δικό του αρχείο καταγραφής γεγονότων αλλά και μεταξύ των αρχείων καταγραφής. Για όλες τις αιτήσεις διατίθεται ένας αριθμός χαρακτηριστικών που περιέχουν πρόσθετες πληροφορίες σχετικά με την αίτηση και τις σχετικές προσφορές . Στα πλαίσια της διπλωματικής όμως χρησιμοποιήθηκαν μόνο οι δραστηριότητες (activities).

Το αρχείο καταγραφής γεγονότων για τη διαδικασία της αίτησης (Application Event Log) περιέχει όλα τα γεγονότα που σχετίζονται με τη διαδικασία αίτησης δανείου, καθώς και πρόσθετες πληροφορίες για αυτήν. Μέσα στο αρχείο καταγραφής, μπορούν να διακριθούν διαφορετικοί τύποι συμβάντων που προκύπτουν από διαφορετικές υποδιεργασίες ή αιτήσεις: Τα συμβάντα τύπου A (επισημαίνονται με το πρόθεμα A στην περιγραφή συμβάντος) αναφέρονται στις διεργασίες που έχουν να κάνουν με τις αιτήσεις, τα συμβάντα τύπου O αναφέρονται στις διεργασίες που έχουν να κάνουν με τις προσφορές ενώ τα συμβάντα τύπου W αναφέρονται σε δραστηριότητες ροής εργασίας (workflow). Συνολικά, υπάρχουν 1.202.267 συμβάντα που αφορούν 31.509 αιτήσεις δανείου. Για αυτές τις εφαρμογές, δημιουργήθηκαν συνολικά 42.995 προσφορές. Για κάθε περίπτωση, υπάρχουν 15 επιπλέον χαρακτηριστικά εκτός από το μοναδικό αναγνωριστικό περίπτωσης (ID), τη χρονική σήμανση και την περιγραφή του γεγονότος. Τα χαρακτηριστικά περιλαμβάνουν, για παράδειγμα, το ζητούμενο ποσό δανείου (αξία σε νόμισμα ευρώ), το credit score του αιτούντος (ακέραια βαθμολογία), τον λόγο για τον οποίο ζητήθηκε το δάνειο (κατηγορικό δεδομένο) και, τον αριθμό των όρων για μια αίτηση (ακέραιος αριθμός αριθμός).

Στην Εικόνα 24 παρουσιάζεται παράδειγμα των πρώτων 3 σειρών και κάποιων από των σημαντικών στηλών του αρχείου καταγραφής γεγονότων για τις διαδικασία της αίτησης δανείου.

	org:resource	concept:name	EventOrigin	EventID	time:timestamp
0	User_1	A_Create Application	Application	Application_652823628	2016-01-01 09:51:15.304000+00:00
1	User_1	A_Submitted	Application	ApplState_1582051990	2016-01-01 09:51:15.352000+00:00
2	User_1	W_Handle leads	Workflow	Workitem_1298499574	2016-01-01 09:51:15.774000+00:00

Εικόνα 24 Οι πρώτες 3 σειρές του event log για το application dataset

Το αρχείο καταγραφής γεγονότων για τη διαδικασία της προσφοράς (Εικόνα 25) περιέχει όλα τα συμβάντα που σχετίζονται με τη διαδικασία δημιουργίας προσφορών και τον χειρισμό αυτών των αποτελεσμάτων από τις εισερχόμενες αιτήσεις δανείου. Υπάρχουν συνολικά 193.849 συμβάντα καταγεγραμμένα στο αρχείο καταγραφής, που αντιστοιχούν σε 42.995 προσφορές. Το αρχείο αυτό ουσιαστικά είναι ένα υποσύνολο του προηγούμενου. Εκτός από το αναγνωριστικό περίπτωσης, τη χρονική σήμανση και την περιγραφή συμβάντος, υπάρχουν 14 ακόμη χαρακτηριστικά στο αρχείο καταγραφής, που περιγράφουν για παράδειγμα το ποσό που προσφέρθηκε στον αιτούντα και το αρχικό ποσό ανάληψης (και τα δύο ποσά σε νόμισμα ευρώ), τον αριθμό των συμφωνηθέντων όρων αποπληρωμής (ακέραια τιμή), και το μηνιαίο κόστος (ποσό σε ευρώ).

	org:resource	concept:name	EventID	time:timestamp	case:concept:name
0	User_17	O_Create Offer	Offer_247135719	2016-01-02 09:17:05.720000+00:00	Offer_247135719
1	User_17	O_Created	OfferState_124849367	2016-01-02 09:17:08.762000+00:00	Offer_247135719
2	User_17	O_Sent (online only)	OfferState_440662877	2016-01-02 09:19:21.330000+00:00	Offer_247135719

Εικόνα 25 Οι πρώτες 3 σειρές του event log για το offer dataset

Όλα τα αρχεία δεδομένων ήταν σε τυπική μορφή XES. Το XES είναι ένα ακρώνυμο για τη ροή συμβάντων eXtensible και βασίζεται στη μορφή αρχείου XML. Έτσι, παρέχει σαφή δομή, μπορεί εύκολα να αναλυθεί και να δημιουργηθεί, και είναι αρκετά ευέλικτο ώστε να συλλαμβάνει λεπτομερή δεδομένα καταγραφής γεγονότων καθώς και πλούσιες πρόσθετες πληροφορίες διεργασίας. Σε σύγκριση με αρχεία καταγραφής γεγονότων απλού κειμένου, όπως για παράδειγμα το .csv που χρησιμοποιείται ευρέως, το XES παρέχει μια πολύ πιο πλούσια παρουσίαση δεδομένων διεργασιών και πρόσθετων χαρακτηριστικών που μπορούν εύκολα να επεξεργαστούν με τυπικά εργαλεία εξόρυξης διεργασιών.

5.3 Process Mining- Εξόρυξη Διαδικασιών

Η εξόρυξη διεργασιών επιτρέπει την ανακάλυψη, ανάλυση και βελτίωση επιχειρηματικών διαδικασιών χρησιμοποιώντας δεδομένα γεγονότων. Για την καλύτερη κατανόηση των δεδομένων αλλά και για σκοπούς σύγκρισης με αλγορίθμους βαθιάς μηχανικής μάθησης εφάρμοσα διάφορες τεχνικές και αλγόριθμους εξόρυξης διαδικασιών. Για να το κάνω αυτό χρησιμοποίησα τη βιβλιοθήκη PM4PY στην Python. Η Pm4py είναι μια βιβλιοθήκη python ανοιχτού κώδικα που δημιουργήθηκε από το Ινστιτούτο Fraunhofer για Εφαρμοσμένη Τεχνολογία Πληροφοριών για την υποστήριξη της Εξόρυξης Διαδικασιών.

Δεδομένου ότι έχω συμπεριλάβει βασικές εννοιολογικές γνώσεις σχετικά με την εξόρυξη διεργασιών και τη διαμόρφωση δεδομένων και συμβάντων σε προηγούμενη ενότητα, τώρα επικεντρώνομαι στην ανακάλυψη της διαδικασίας (Process Discovery). Ο στόχος δηλαδή είναι να ανακαλύψουμε, πλήρως αυτοματοποιημένα και αλγοριθμικά, ένα μοντέλο διεργασίας που περιγράφει με ακρίβεια τη διαδικασία, δηλαδή, όπως παρατηρείται στα δεδομένα γεγονότων.

Αυτή η ενότητα εξηγεί εν συντομία τι φορμαλισμούς μοντελοποίησης υπάρχουν στο PM4Py ενώ εφαρμόζουμε διαφορετικούς αλγόριθμους ανακάλυψης διαδικασίας.

Η συνολική διαδικασία υποβολής αιτήσεων και χειρισμού δανείων αποτελείται από τις τρία κομμάτια (την αίτηση, την προσφορά και τη ροή εργασιών) που περιέχουν συνολικά 26 δραστηριότητες (Πίνακας 5). Από μια γενική οπτική μέσα από τα αποτελέσματα των miners, η συνολική διαδικασία μπορεί να περιγραφεί ως εξής: η διαδικασία ξεκινά με τη δημιουργία και την υποβολή μιας νέας αίτησης. Μετά από αυτό, εκτελούνται ορισμένα εσωτερικά προπαρασκευαστικά βήματα για την εκτέλεση αυτόματων ελέγχων στην εφαρμογή και τη δημιουργία μιας νέας ροής εργασίας. Σε ορισμένες περιπτώσεις, ζητούνται πρόσθετες πληροφορίες από τον αιτούντα πριν από την περαιτέρω επεξεργασία της αίτησης. Στη συνέχεια, επικυρώνεται η πλήρης αίτηση, η οποία μπορεί να οδηγήσει σε τρία αποτελέσματα: η αίτηση μπορεί αμέσως να απορριφθεί (π.χ. λόγω τυπικών ζητημάτων), να γίνει απευθείας αποδεκτή (που ονομάζεται "συντομευμένη ολοκλήρωση") ή να επικυρωθεί με περισσότερες λεπτομέρειες. Στην πρώτη περίπτωση, η εφαρμογή θα ακυρωθεί και η διαδικασία θα τερματιστεί. Στη δεύτερη περίπτωση, δημιουργείται μια προσφορά, αποστέλλεται στον πελάτη (μόνο μέσω ταχυδρομείου ή μέσω διαδικτύου) και συζητείται μαζί τους στο τηλέφωνο. Οι δραστηριότητες προσφοράς είναι συνολικά 8. Στην τρίτη περίπτωση, πραγματοποιείται σε βάθος ανάλυση, η οποία περιλαμβάνει εντοπισμό απάτης και προαιρετικά αιτήματα για πιθανώς ελλιπή αρχεία εφαρμογών. Ως αποτέλεσμα, μια προσφορά μπορεί να δημιουργηθεί όταν η επικύρωση περάσει ή η εφαρμογή μπορεί να απορριφθεί όταν η επικύρωση αποτύχει.

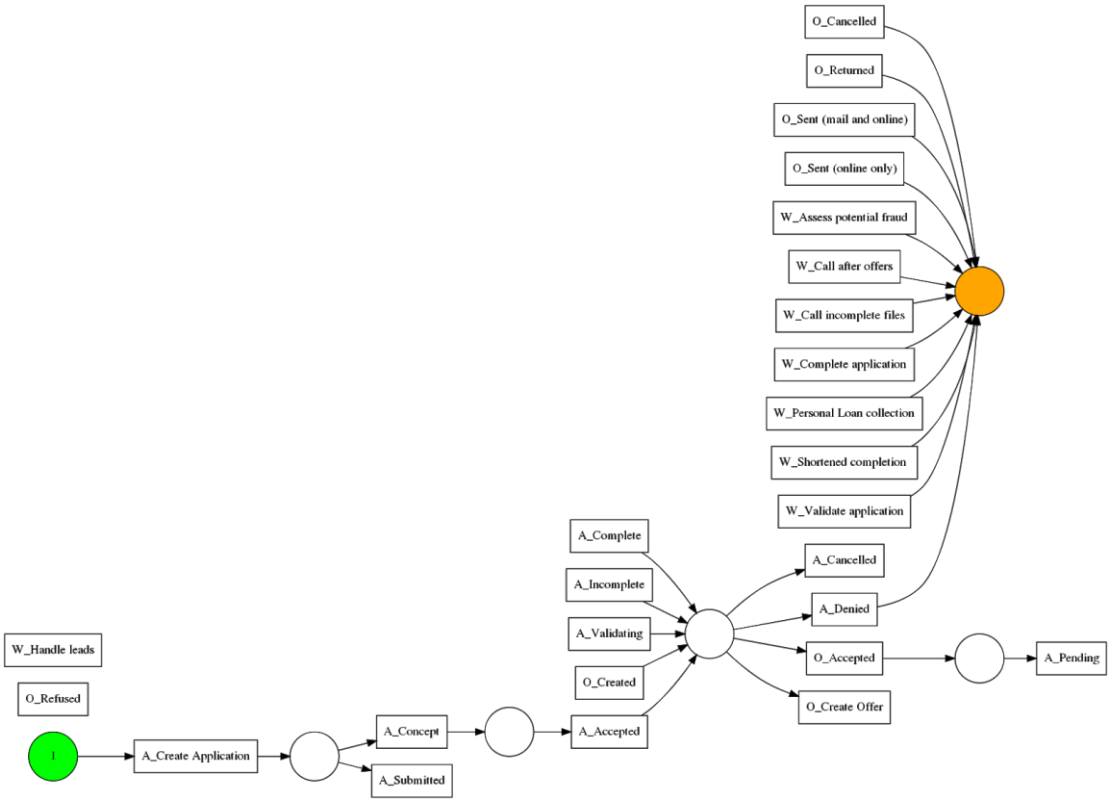
Πίνακας 5 Οι δραστηριότητες του event log

Κατηγορία Δραστηριότητας	Δραστηριότητα
'A' Δραστηριότητες σχετικές με την αίτηση (application) δανείου	A_Create_Application
	A_Submited
	A_Concept
	A_Accepted
	A_Complete
	A_Validating
	A_Incomplete
	A_Pending
	A_Denied
	A_Cancelled
'O' Δραστηριότητες σχετικές με την προσφορά (offer) δανείου	O_Create Offer
	O_Created
	O_Sent (mail and online)
	O_Sent (online only)
	O_Returned
	O_Accepted
	O_Refused
	O_Cancelled
'W Δραστηριότητες σχετικές με το Workflow	W_Handle leads
	W_Complete application
	W_Call after offers
	W_Validate application
	W_Call Incomplete files
	W_Assess potential fraud
	W_PersonalLoan collection
	W_Shortened completion

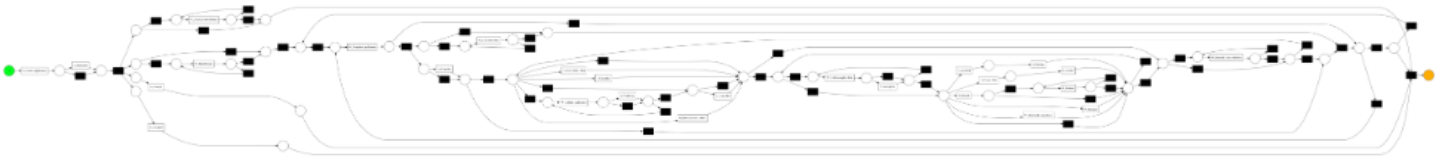
Στις επόμενες υποενότητες θα παρουσιαστούν τα αποτελέσματα 3 αλγορίθμων για ανακάλυψη διαδικασιών για τα δύο σύνολα δεδομένων. Οι τρεις αυτοί αλγόριθμοι είναι:

1. Alpha Miner
2. Inductive Miner
3. Heuristic Miner

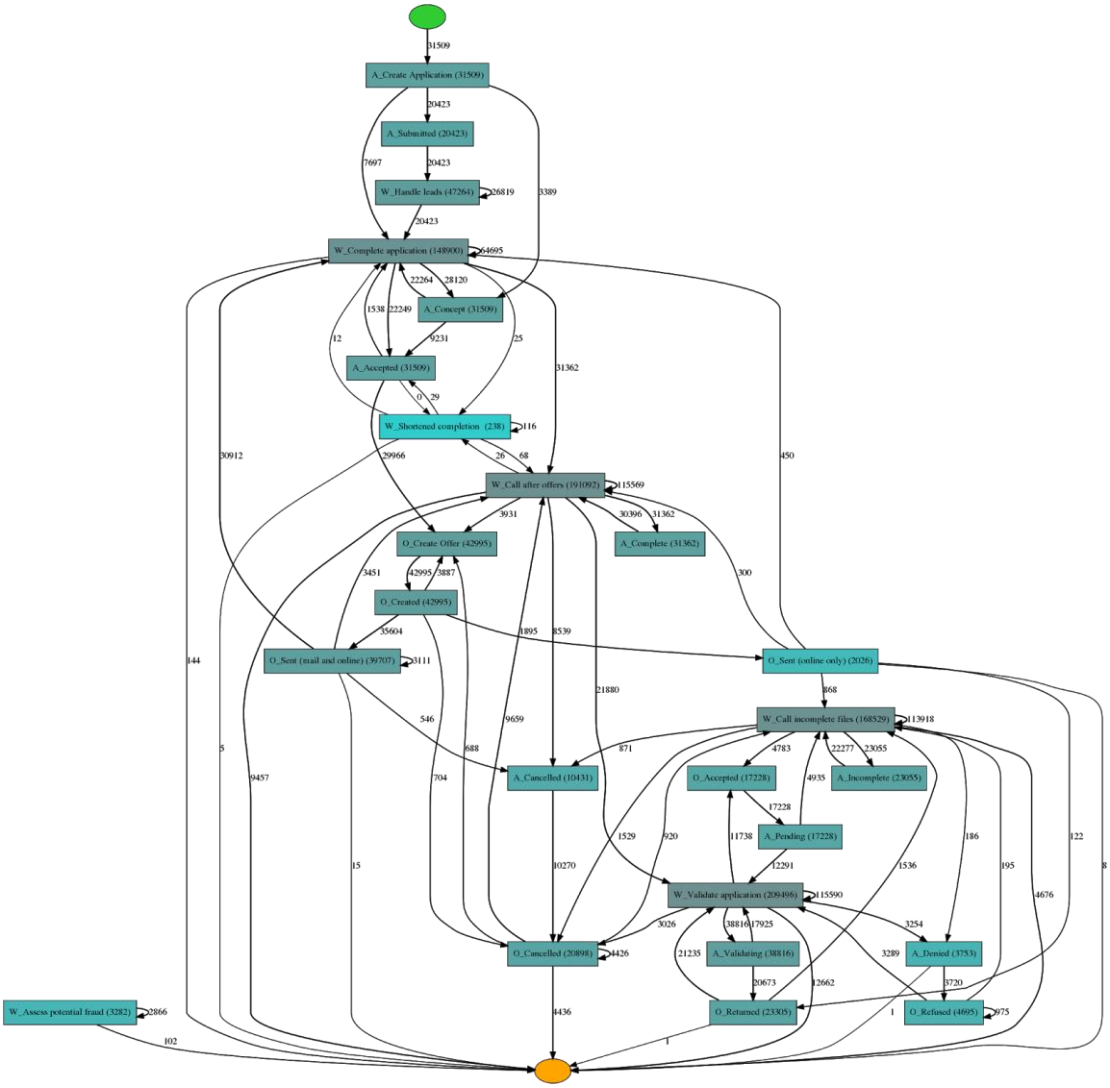
5.3.1 Ανακάλυψη Διαδικασιών – Δεδομένα Αίτησης Δανείου



Εικόνα 26 Alpha Miner - Application Dataset



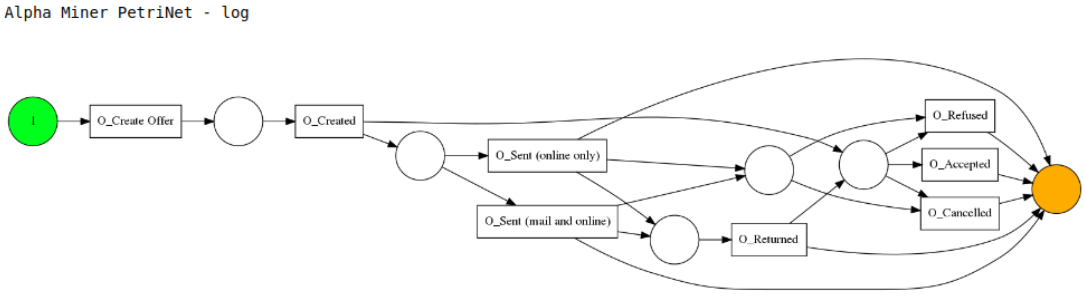
Εικόνα 27 Inductive Miner - Application Dataset



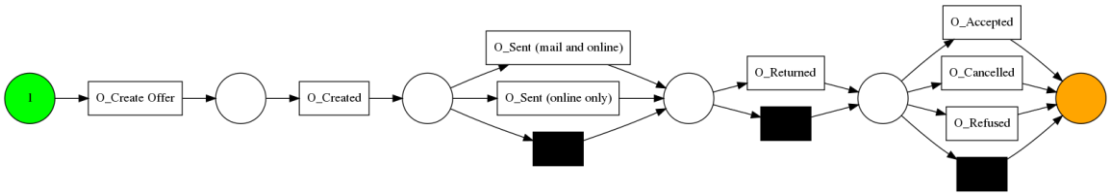
Εικόνα 28 Heuristic Miner - Application Dataset

Παρατηρούμε λοιπόν όπως ήταν αναμενόμενο ότι τα συγκεκριμένα μοντέλα διεργασιών που μοιάζουν με σπαγγέτι και αποτελούν μέρος του χαμηλότερου επιπέδου μοντέλων διεργασιών είναι δύσκολο να ερμηνευτούν και να κατανοηθούν από τα ανθρώπινα όντα.

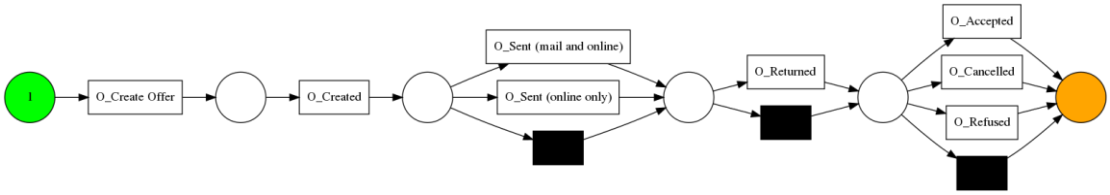
5.3.2 Ανακάλυψη Δαδικασιών – Δεδομένα Προσφοράς Δανείου



Εικόνα 29 Alpha Miner - Offer Dataset



Εικόνα 30 Inductive Miner - Offer Dataset



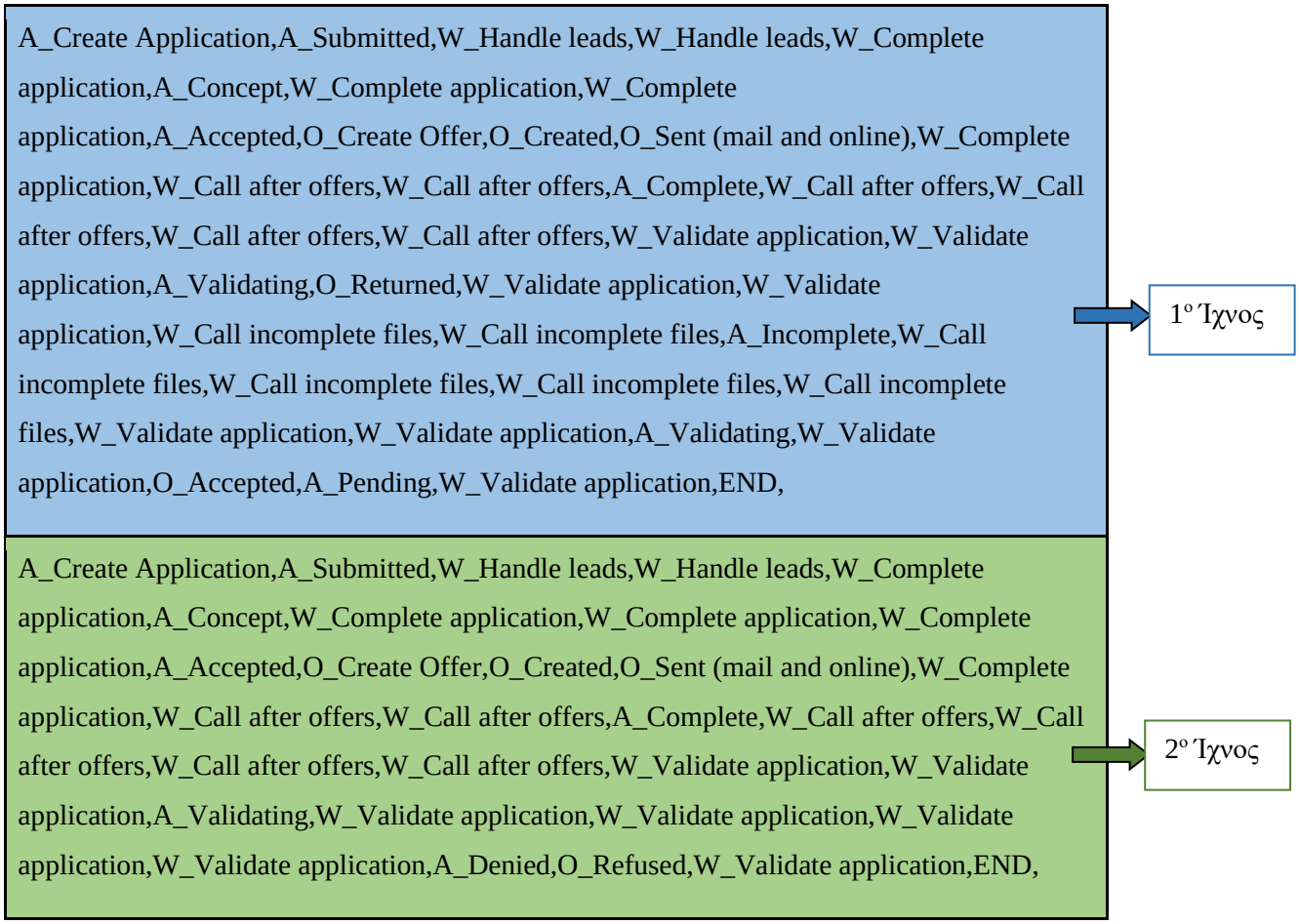
Εικόνα 31 Heuristic Miner - Offer Dataset

5.4 Προεπεξεργασία Δεδομένων

5.4.1 Απομόνωση των Ιχνών

Το πρώτο βήμα της προ-επεξεργασίας των δεδομένων πριν να δοθούν ως είσοδος στο νευρωνικό δίκτυο, είναι η απομόνωση των ιχνών (traces) και η αποθήκευση τους σε ξεχωριστό αρχείο. Με τη βοήθεια ενός python script θα δημιουργηθεί ένα αρχείο κειμένου που σε κάθε γραμμή θα έχει τις δραστηριότητες ενός trace διατηρώντας την σειρά εμφάνισης τους. Στο τέλος της κάθε γραμμής τοποθετείται και η λέξη END για να σηματοδοτεί το που τελειώνει το κάθε trace.

Στην Εικόνα 32 φαίνονται τα δύο πρώτα ίχνη στα δεδομένα της αίτησης δανείου (Application Dataset).



Εικόνα 32 Τα 2 πρώτα ίχνη του application dataset

5.4.2 Δημιουργία Λεξικού Ακεραίων

Στη συνέχεια δημιούργησα το λεξικό και αντιστοίχησα τις δραστηριότητες με ένα ακέραιο αριθμό την κάθε μία. Η αντιστοίχιση που προέκυψε από τον tokenizer της python φαίνεται πιο κάτω.

Για το αρχείο δεδομένων της αίτησης (Πίνακας 6):

Πίνακας 6 Λεξικό ακεραίων (application dataset)

Όνομα δραστηριότητας	Ακέραιος	Όνομα δραστηριότητας	Ακέραιος
w_validateapplication	1	**o_returned**	15
w_callafteroffers	2	**a_incomplete**	16
w_callincompletefiles	3	**o_cancelled**	17
w_completeapplication	4	**a_submitted**	18
w_handleleads	5	**o_accepted**	19
o_createoffer	6	**a_pending**	20
o_created	7	**a_cancelled**	21
o_sent(mailandonline)	8	**o_refused**	22
a_validating	9	**a_denied**	23
a_createapplication	10	**w_assesspotentialfraud**	24
a_concept	11	**o_sent(onlineonly)**	25
a_accepted	12	**w_shortenedcompletion**	26
end	13	**w_personalloancollection**	27
a_complete	14		

Για το αρχείο δεδομένων της προσφοράς (Πίνακας 7):

Πίνακας 7 Λεξικό ακεραίων (offer dataset)

Όνομα δραστηριότητας	Ακέραιος
o_createoffer	1
o_created	2
end	3
o_sent(mailandonline)	4
o_returned	5
o_cancelled	6
o_accepted	7
o_refused	8
o_sent(onlineonly)	9

5.4.3 *Κωδικοποίηση One-Hot*

Το επόμενο βήμα είναι η κωδικοποίηση one-hot encoding να εφαρμοστεί στην ακέραια αναπαράσταση.

Ένα παράδειγμα του πως εφαρμόζεται η κωδικοποίηση αυτή στο αρχείο δεδομένων προσφοράς φαίνεται πιο κάτω (Εικόνα 33):

Όνομα δραστηριότητας	Ακέραιος
o_createoffer	1
o_created	2
end	3

Όνομα δραστηριότητας	o_create offer	o_created	end	o_sent(mail and online)	o_returned	o_cancelled	o_accepted	o_refused	o_sent (online only)
o_createoffer (1)	**1**	0	0	0	0	0	0	0	0
O_created (2)	0	**1**	0	0	0	0	0	0	0
End (3)	0	0	**1**	0	0	0	0	0	0

Εικόνα 33 Παράδειγμα one-hot encoding

Μετά την κωδικοποίηση όλων των δραστηριοτήτων που υπάρχουν στο αρχείο κειμένου έχουμε ως αποτέλεσμα έναν πίνακα που αποτελείται από ένα αριθμό διανυσμάτων ίσο με τον αριθμό των δραστηριοτήτων στο αρχείο και που το κάθε ένα έχει αριθμό στηλών ίσο με τον αριθμό των μοναδικών δραστηριοτήτων (δραστηριότητες και σήμανση END) σε κάθε αρχείο (27 για το αρχείο αίτησης και 9 για το αρχείο της προσφοράς). Ο πίνακας αυτός είναι της πιο κάτω μορφής:

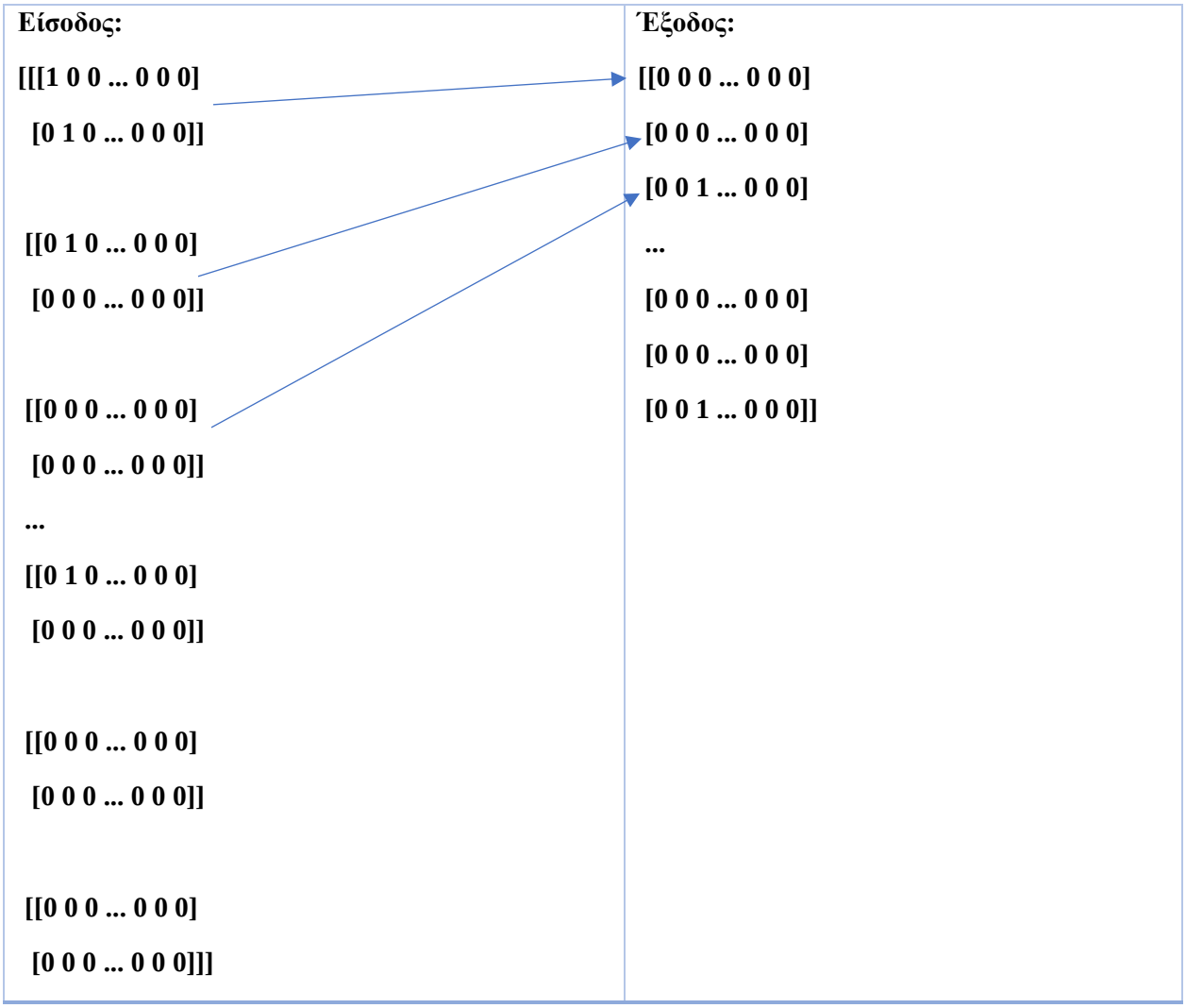
```
[[1 0 0 ... 0 0 0]
 [0 1 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 1 ... 0 0 0]]
```

5.4.4 *Δημιουργία διανυσμάτων Εισόδου- Εξόδου*

Στη συνέχεια πήρα ανά ζευγάρια τα activities των traces έτσι ώστε να δημιουργήσω τα διανύσματα εισόδου και εξόδου. Υπάρχει η δυνατότητα να γίνει πειραματισμός με πολλούς

συνδυασμούς για παράδειγμα μια δραστηριότητα ως είσοδος και μία ως έξοδος, 2 ή περισσότερες δραστηριότητες ως είσοδος και μία ή περισσότερες ως έξοδος. Στο αρχείο της προσφοράς με αυτό το μοτίβο δημιουργούνται 149860 ζευγάρια ενώ στης αίτησης 1048234 .

Στο παράδειγμα που ακολουθεί (Εικόνα 34) δείχνω τη δημιουργία ζευγαριών που έχουν ως είσοδο 2 συνεχόμενες δραστηριότητες και η αμέσως επόμενη στη σειρά είναι η έξοδος.



Εικόνα 34 Παράδειγμα διανυσμάτων εισόδου-εξόδου

5.5 Πειράματα για εύρεση βέλτιστων υπερ παραμέτρων

Κατά το σχεδιασμό της αρχιτεκτονικής ενός νευρωνικού δικτύου, υπάρχει μια ποικιλία παραμέτρων που μπορούν να συντονιστούν (parameter tuning). Είναι πράγματι μια τέχνη από μόνη της η εύρεση του σωστού συνδυασμού για αυτές τις παραμέτρους για την επίτευξη της υψηλότερης ακρίβειας (accuracy) και της χαμηλότερης απώλειας (loss). Στα πλαίσια της διπλωματικής μου χρησιμοποίησα το Talos για βελτιστοποίηση των υπερ-παραμέτρων του νευρωνικού μου δικτύου.

Ακριβώς όπως άλλες γνωστές τεχνικές όπως το GridSearchCV για βελτιστοποίηση υπερ-παραμέτρων σε μοντέλα scikit-learning, όπως Δέντρα Απόφασης / Τυχαίο Δάσος και Μηχανές Διανυσμάτων Υποστήριξης(Support Vector Machine), το Talos μπορεί να εφαρμοστεί σε μοντέλα Keras. Το Talos λειτουργεί παρόμοια με το GridSearchCV, δοκιμάζοντας όλους τους πιθανούς συνδυασμούς αυτών των παραμέτρων που έχουμε εισαγάγει και επιλέγει το καλύτερο μοντέλο, με βάση την παράμετρο που ζητήσαμε να βελτιστοποιήσει ή να μειώσει.

Οι παράμετροι: αλγόριθμος βελτιστοποίησης (optimizer), συνάρτηση ενεργοποίησης (activation function) και συνάρτηση απώλειας (loss function) καθορίστηκαν όπως αναφέρθηκε στο προηγούμενο κεφάλαιο με βάση των τύπο των δεδομένων μου.

Κατά τη διάρκεια του συντονισμού των υπερ-παραμέτρων όμως χρειάστηκε να ληφθούν αποφάσεις σχετικά με την επιλογή παραμέτρων όπως ο κατάλληλος αριθμός επιπέδων (layers), ο κατάλληλος αριθμός νευρώνων (neurons), το ποσοστό του περιορισμού ενεργοποίησης (dropout) και του Batch Size.

Τα πειράματα επαναλήφθηκαν για όλους του συνδυασμούς προβλέψεων που έκανα. Δηλαδή για διαφορετικό αριθμό δραστηριοτήτων που έδινα για είσοδο και έξοδο.

Αξίζει να σημειωθεί ότι έγινε πειραματισμός και με περισσότερα κρυφά επίπεδα αλλά κρίθηκε πως με ένα κρυφό επίπεδο γινόταν επίτευξη των καλύτερων αποτελεσμάτων.

5.5.1 Πειράματα για 2 δραστηριότητες ως είσοδο και 1 ως έξοδο - Application Dataset

Στον ακόλουθο πίνακα (Πίνακας 8) φαίνεται πως όρισα όλες τις παραμέτρους που ήθελα να δοκιμάσει το μοντέλο μου.

Πίνακας 8 Παραμέτροι που θα δοκιμάσει το μοντέλο- 2 δραστηριότητες ως είσοδο και 1 ως έξοδο - Application Dataset

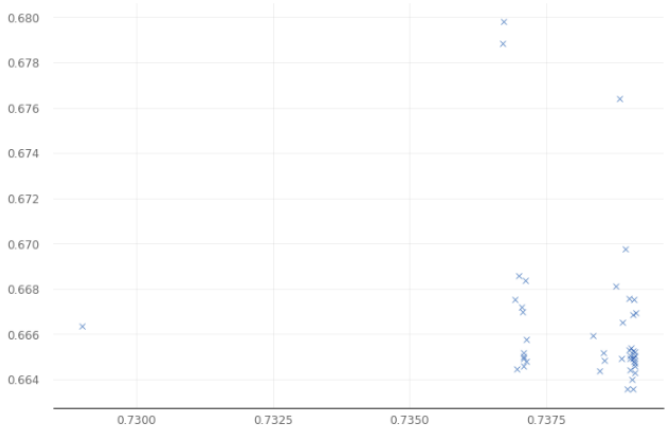
Αριθμός Νευρώνων (neurons)	**[20,40,50,60,80]**
Ποσοστό του περιορισμού ενεργοποίησης (dropout)	**[0,0.1,0.5]**
Batch Size	**[64, 128, 256]**

Στον επόμενο πίνακα (Πίνακας 9) παρουσιάζεται ένα δείγμα από τα πρώτα 10 αποτελέσματα των συνολικά 45 πειραμάτων που έγιναν για 15 εποχές στις οποίες παρατηρήθηκε ότι ήταν αρκετές για επίτευξη βέλτιστων αποτελεσμάτων. Ο πίνακας είναι ταξινομημένος σε φθίνουσα σειρά με βάση το validation accuracy.

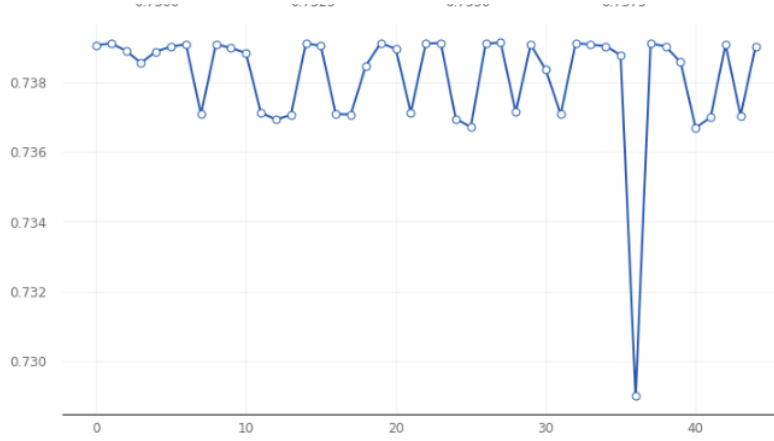
Πίνακας 9 Αποτελέσματα πειραμάτων- 2 δραστηριότητες ως είσοδο και 1 ως έξοδο - Application Dataset

start	end	duration	loss	accuracy	Val loss	Val accuracy	Batch size	dropout	First neuron
06/06/21-145111	06/06/21-145536	264.591086	0.68688	0.736217	0.666957	0.739136	128	0.5	50
06/06/21-141355	06/06/21-142019	384.534122	0.667284	0.737222	0.665223	0.739127	128	0	80
06/06/21-143131	06/06/21-143656	324.603059	0.669333	0.73758	0.664594	0.739127	128	0.1	60
06/06/21-151237	06/06/21-151521	163.362004	0.666788	0.737822	0.665036	0.739127	256	0	50
06/06/21-121657	06/06/21-122321	384.542545	0.666911	0.737486	0.664701	0.739117	64	0	40
06/06/21-134814	06/06/21-135639	504.555332	0.677087	0.73697	0.664293	0.739117	64	0.5	80
06/06/21-153009	06/06/21-153301	172.573919	0.671009	0.737116	0.664796	0.739112	256	0.1	50
06/06/21-142706	06/06/21-143131	264.679841	0.669826	0.737325	0.66498	0.739108	128	0.1	50
06/06/21-144646	06/06/21-145111	264.633119	0.695355	0.735175	0.667522	0.739108	128	0.5	40
06/06/21-125111	06/06/21-125736	384.567913	0.670021	0.737734	0.664902	0.739089	64	0.1	40

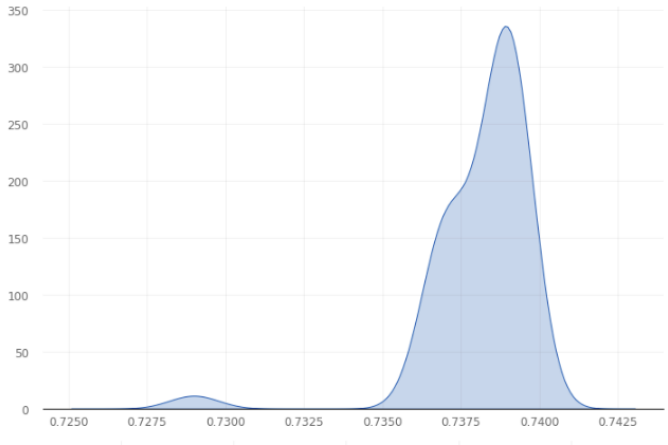
Για καλύτερη οπτικοποίηση των αποτελεσμάτων των πειραμάτων για τις υπερ παραμέτρους παράχθηκαν διάφορες γραφικές οι οποίες παρατίθενται πιο κάτω.



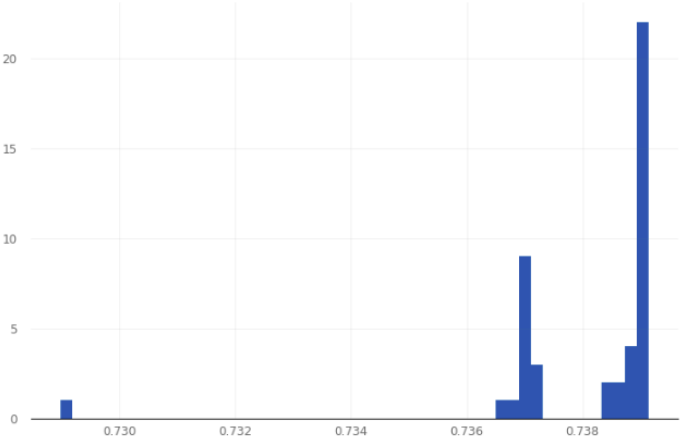
Εικόνα 35 Validation Accuracy vs Validation Loss



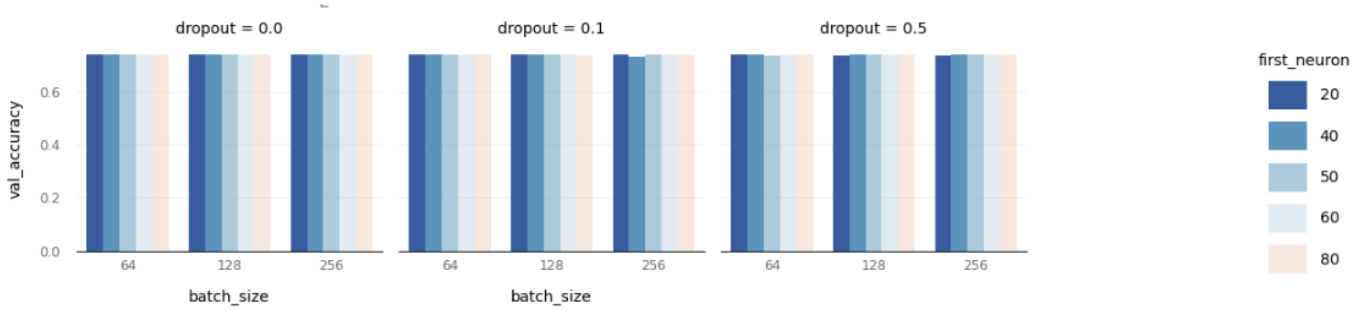
Εικόνα 36 Validation Accuracy of all the experiments



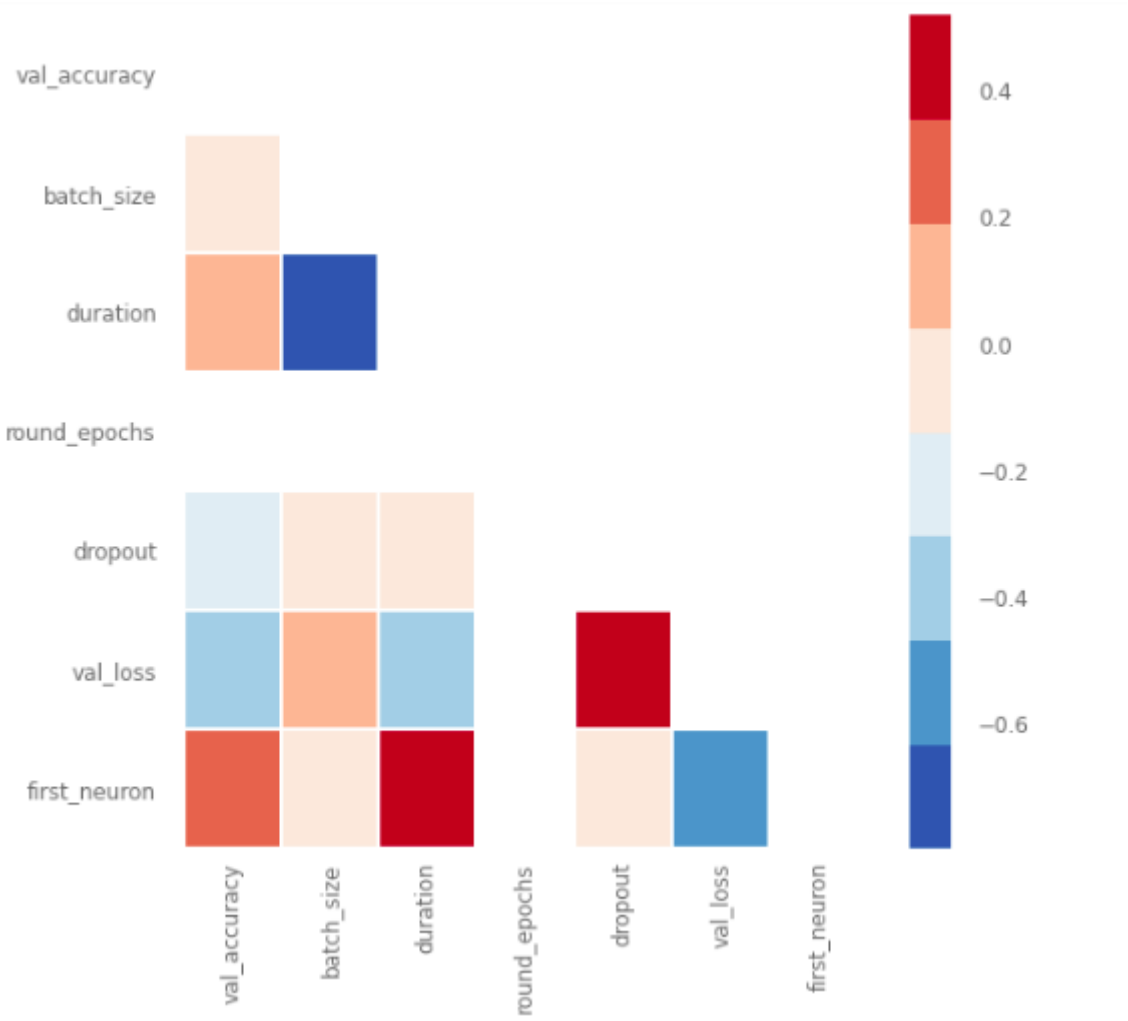
Εικόνα 37 Validation Accuracy of all the experiments - KDE Plot



Εικόνα 38 Ιστόγραμμα Validation Accuracy



Εικόνα 39 Bar Grind



Εικόνα 40 Heatmap Correlation

Στην Εικόνα 35 παρουσιάζονται το validation accuracy έναντι του validation loss για όλα τα πειράματα που έγιναν ενώ στην Εικόνα 36 παρουσιάζεται το validation accuracy για όλα τα πειράματα. Παρόμοια στις αμέσως επόμενες εικόνες (Εικόνα 37,Εικόνα 38) έχουμε για το validation accuracy το ιστόγραμμα και το γράφημα KDA αντίστοιχα. Τέλος στις τελευταίες εικόνες (Εικόνα 39, Εικόνα 40) παρουσιάζονται σε grid και heatmap αντίστοιχα το validation accuracy σε σχέση με τις διάφορες υπερ-παραμέτρους.

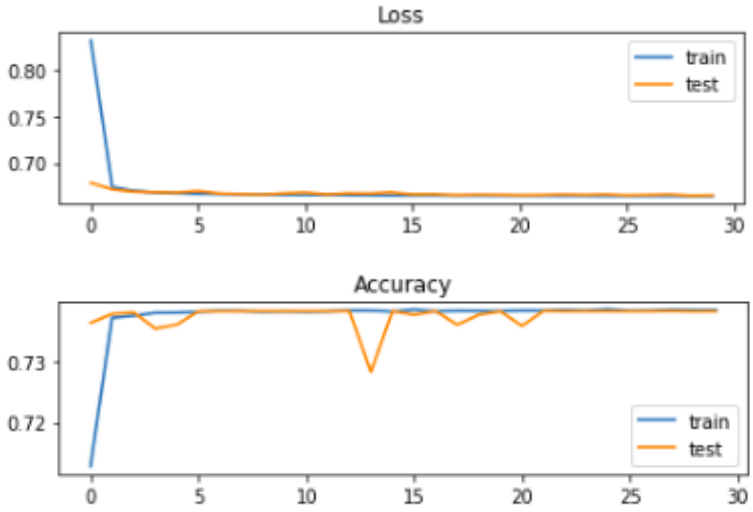
Για τα δεδομένα της αίτησης δανείου και της πρόβλεψης με 2 εισόδους επιλέχθηκαν τελικά οι πιο κάτω υπερ παράμετροι (Εικόνα 41):

Αριθμός Νευρώνων (neurons)	**50**
Ποσοστό του περιορισμού ενεργοποίησης (dropout)	**0**
Batch Size	**256**

Εικόνα 41 Τελική επιλογή υπερ παραμέτρων- 2 δραστηριότητες ως είσοδο και 1 ως έξοδο - Application Dataset

Ο λόγος που έγινε οι πιο πάνω επιλογή είναι γιατί οι συγκεκριμένες παράμετροι είχαν από τα πιο ψηλά accuracy και χαμηλό loss. Υπάρχουν βέβαια επιλογές με ελάχιστα καλύτερο συνδυασμό των δύο μετρικών οι οποίες όμως δεν προτιμήθηκαν αφού το χρονικό διάστημα που χρειάζονταν τα δεδομένα για να γίνουν train ήταν πολύ υψηλότερο (διπλάσιο και τριπλάσιο ορισμένες φορές). Έτσι, τελικά προτιμήθηκαν οι πιο πάνω οι οποίες είχαν εξαιρετικά accuracy και loss και ταυτόχρονα χαμηλό χρόνο εκπαίδευσης των δεδομένων.

Στα επόμενα γραφήματα φαίνεται η εξέλιξη των loss και accuracy για τις επιλεγμένες υπερ-παραμέτρους κατά τη πάροδο της εκπαίδευσης 30 εποχών. Το πρώτο γράφημα δείχνει την απώλεια εγκάρσιας εντροπίας για κάθε εποχή για το σύνολο δεδομένων της εκπαίδευσης (μπλε) και της δοκιμής (πορτοκαλί), και η κάτω γραφική παράσταση δείχνει την ακρίβεια (accuracy) έναντι των εποχών.



Εικόνα 42 Απώλεια εγκάρσιας εντρόπιας και ακρίβεια έναντι των εποχών

Παρατηρούμε ότι, τα γραφήματα δείχνουν ότι το μοντέλο φαίνεται να συγκλίνει. Τα γραφήματα τόσο για την απώλεια εγκάρσιας εντροπίας όσο και για ακρίβεια δείχνουν καλή συμπεριφορά σύγκλισης η οποία ταιριάζει με τα θεωρητικά επιθυμητά αποτελέσματα. Το μοντέλο μπορεί να είναι καλά διαμορφωμένο δεδομένου ότι δεν υπάρχει ένδειξη για underfiiting.

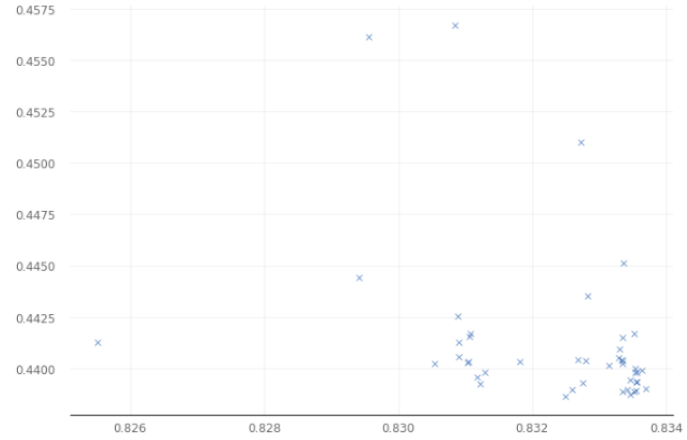
5.5.2 Πειράματα για 3 δραστηριότητες ως είσοδο και 1 ως έξοδο - Application Dataset

Στον επόμενο πίνακα παρουσιάζεται ένα δείγμα από τα πρώτα 10 αποτελέσματα των συνολικά 45 πειραμάτων που έγιναν για 15 εποχές στις οποίες παρατηρήθηκε ότι ήταν αρκετές για επίτευξη βέλτιστων αποτελεσμάτων. Ο Πίνακας 10 είναι ταξινομημένος σε φθίνουσα σειρά με βάση το validation accuracy.

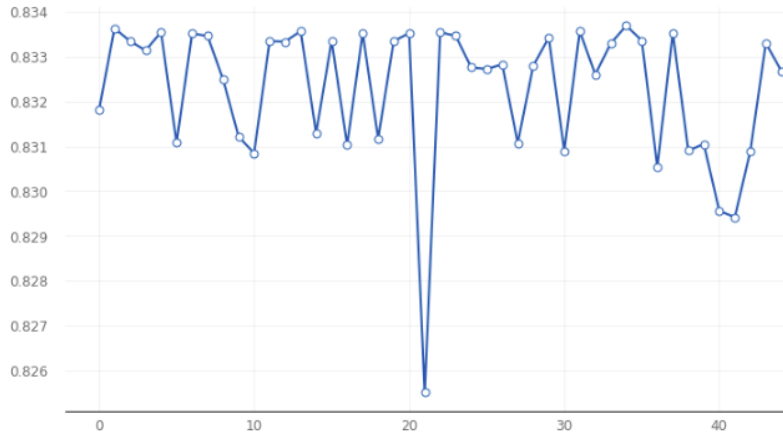
Πίνακας 10 Αποτελέσματα πειραμάτων- 3 δραστηριότητες ως είσοδο και 1 ως έξοδο - Application Dataset

start	end	duration	loss	accuracy	Val loss	Val accuracy	Batch size	dropout	First neuron
06/07/21-135504	06/07/21-140129	384.530030965805	0.440686225891113	0.831791937351227	0.438999325037003	0.833688616752625	256	0	80
06/07/21-104957	06/07/21-105625	388.17768406868	0.440161734819412	0.832350075244904	0.439901530742645	0.833630084991455	64	0	40
06/07/21-121654	06/07/21-122419	444.563494920731	0.454281806945801	0.830619156360626	0.439817488193512	0.833566665649414	64	0.5	60
06/07/21-134457	06/07/21-134734	156.996565818787	0.440741747617722	0.832396090003067	0.439340531826019	0.833561778068543	256	0	40
06/07/21-111216	06/07/21-112041	504.506721973419	0.440204411745071	0.832582116127014	0.438882470130 92	0.833542287349701	64	0	80
06/07/21-130305	06/07/21-130722	257.465741157532	0.443767905235 29	0.831852555274963	0.439326316118 24	0.833542287349701	128	0.1	50
06/07/21-140610	06/07/21-140917	186.521287202835	0.445192366838455	0.831457436084747	0.439982563257217	0.833532512187958	256	0.1	50
06/07/21-124021	06/07/21-124438	257.78209400177	0.440266191959381	0.832456707954407	0.439813971571947	0.833527624607086	128	0	50
06/07/21-125552	06/07/21-125917	204.498727083206	0.452745169401169	0.830497860908508	0.441685020923615	0.833522737026215	128	0.1	20
06/07/21-112641	06/07/21-113305	384.530819416046	0.443732559680939	0.831794023513794	0.438855022192001	0.833513021469116	64	0.1	40

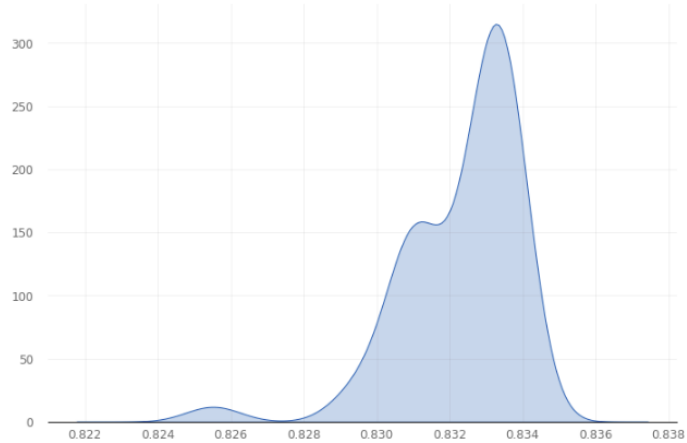
Για καλύτερη οπτικοποίηση των αποτελεσμάτων των πειραμάτων για τις υπερ παραμέτρους παράχθηκαν διάφορες γραφικές οι οποίες παρατίθενται πιο κάτω.



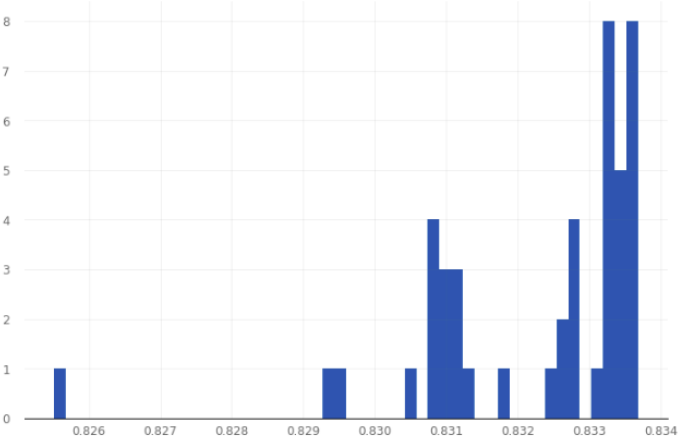
Εικόνα 43 Validation Accuracy vs Validation Loss



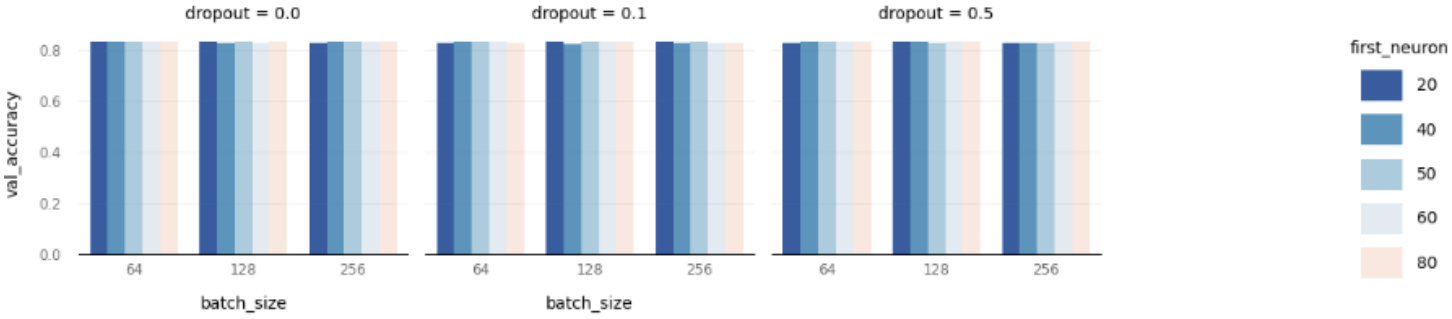
Εικόνα 44 Validation Accuracy of all the experiments



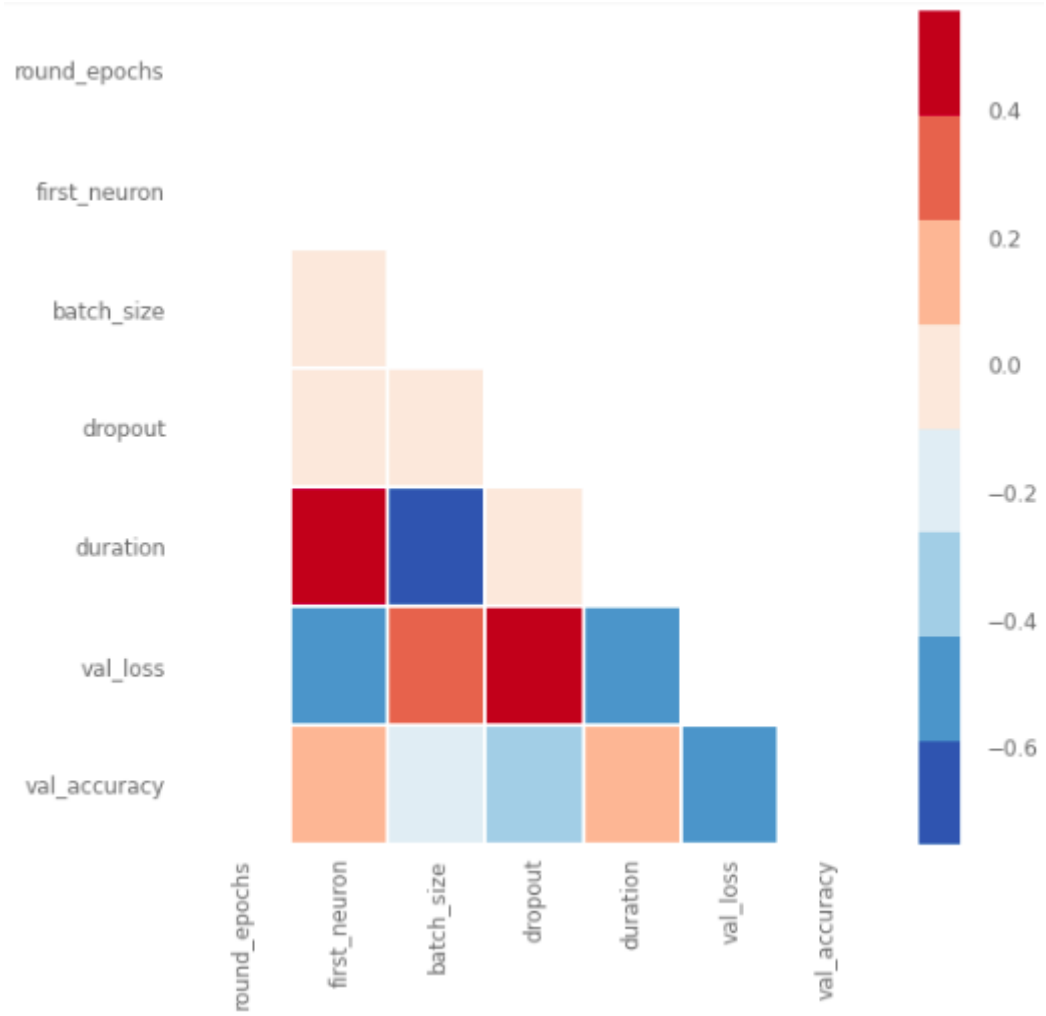
Εικόνα 45 Validation Accuracy of all the experiments - KDE Plot



Εικόνα 46 Ιστόγραμμα Validation Accuracy



Εικόνα 47 Bar Grind



Εικόνα 48 Heatmap Correlation

Για τα δεδομένα της αίτησης δανείου για 3 εισόδους και 1 έξοδο ως πρόβλεψη επιλέχθηκαν τελικά οι πιο κάτω υπερ παράμετροι (Πίνακας 11):

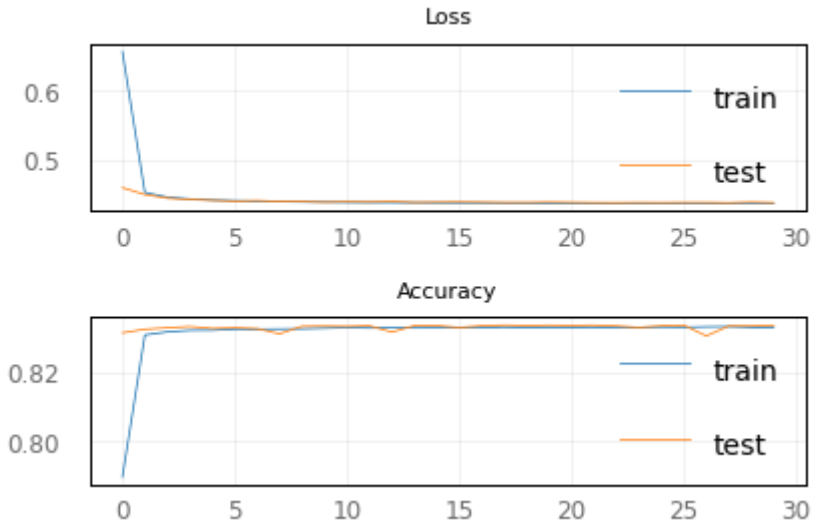
Πίνακας 11 Τελική επιλογή υπερ παραμέτρων-3 δραστηριότητες ως είσοδο και 1 ως έξοδο - Application Dataset

Αριθμός Νευρώνων (neurons)	**40**
Ποσοστό του περιορισμού ενεργοποίησης (dropout)	**0**
Batch Size	**256**

Ο λόγος που έγινε οι πιο πάνω επιλογή είναι γιατί οι συγκεκριμένες παράμετροι είχαν από τα πιο ψηλά accuracy και χαμηλό loss. Υπάρχουν βέβαια επιλογές με ελάχιστα καλύτερο συνδυασμό των δύο μετρικών οι οποίες όμως δεν προτιμήθηκαν αφού το χρονικό διάστημα που χρειάζονταν τα δεδομένα για να γίνουν train ήταν πολύ υψηλότερο (διπλάσιο και τριπλάσιο

ορισμένες φορές). Έτσι, τελικά προτιμήθηκαν οι πιο πάνω οι οποίες είχαν εξαιρετικά accuracy και loss και ταυτόχρονα χαμηλό χρόνο εκπαίδευσης των δεδομένων.

Στα επόμενα γραφήματα (Εικόνα 49) φαίνεται η εξέλιξη των loss και accuracy για τις επιλεγμένες υπερ παραμέτρους κατα τη πάροδο της εκπαίδευσης 30 εποχών. Το πρώτο γράφημα δείχνει την απώλεια εγκάρσιας εντροπίας για κάθε εποχή για το σύνολο δεδομένων της εκπαίδευσης (μπλε) και της δοκιμής (πορτοκαλί), και η κάτω γραφική παράσταση δείχνει την ακρίβεια (accuracy) έναντι των εποχών.



Εικόνα 49 Απώλεια εγκάρσιας εντρόπιας και ακρίβεια έναντι των εποχών

Παρατηρούμε ότι, τα γραφήματα δείχνουν ότι το μοντέλο φαίνεται να συγκλίνει. Τα γραφήματα τόσο για την απώλεια εγκάρσιας εντροπίας όσο και για ακρίβεια δείχνουν καλή συμπεριφορά σύγκλισης η οποία ταιριάζει με τα θεωρητικά επιθυμητά αποτελέσματα. Το μοντέλο μπορεί να είναι καλά διαμορφωμένο δεδομένου ότι δεν υπάρχει ένδειξη για underfiiting.

5.5.3 Πειράματα για 2 δραστηριότητες ως είσοδο και 1 ως έξοδο - Offer Dataset

Στον ακόλουθο πίνακα (Πίνακας 12) φαίνεται πως όρισα όλες τις παραμέτρους που ήθελα να δοκιμάσει το μοντέλο μου.

Πίνακας 12 Παραμέτροι που θα δοκιμάσει το μοντέλο- 2 δραστηριότητες ως είσοδο και 1 ως έξοδο - Offer Dataset

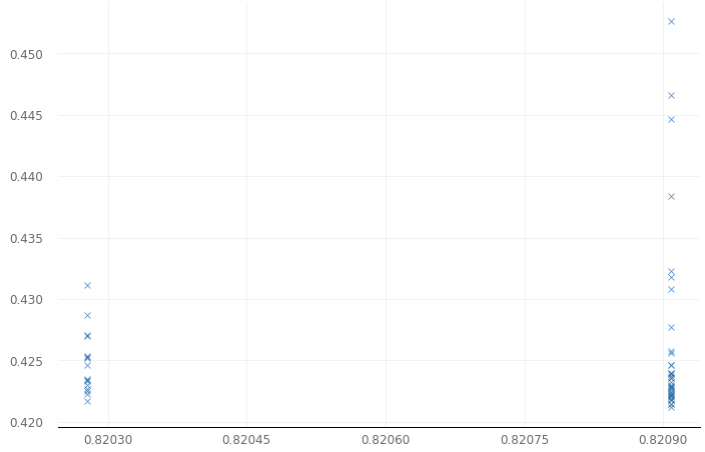
Αριθμός Νευρώνων (neurons)	**[5,10,15,20,25,30]**
Ποσοστό του περιορισμού ενεργοποίησης (dropout)	**[0,0.1,0.5]**
Batch Size	**[32, 64, 128, 256]**

Στον επόμενο πίνακα (Πίνακας 13) παρουσιάζεται ένα δείγμα από τα πρώτα 10 αποτελέσματα των συνολικά 54 πειραμάτων που έγιναν για 15 εποχές στις οποίες παρατηρήθηκε ότι ήταν αρκετές για επίτευξη βέλτιστων αποτελεσμάτων. Ο πίνακας είναι ταξινομημένος σε φθίνουσα σειρά με βάση το validation accuracy.

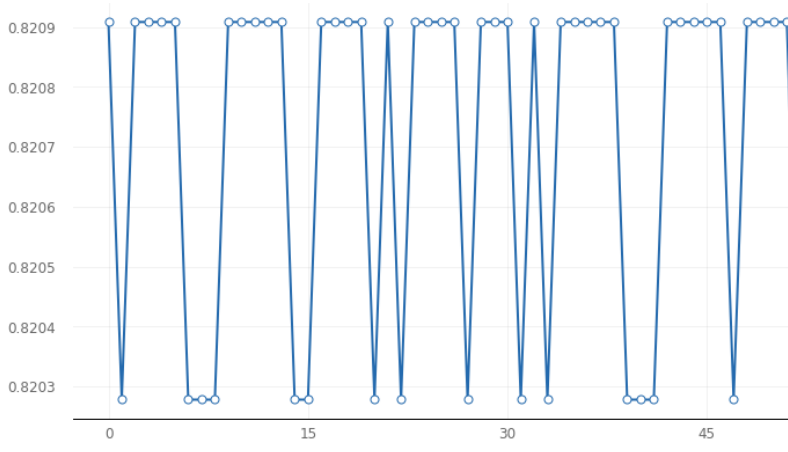
Πίνακας 13 Αποτελέσματα πειραμάτων- 2 δραστηριότητες ως είσοδο και 1 ως έξοδο - Offer Dataset

start	end	duration	loss	accuracy	Val loss	Val accuracy	Batch size	dropout	First neuron
06/07/21-154851	06/07/21-154934	43.5840883255005	0.434872180223465	0.81261819601059	0.422419667243957	0.820908248424 53	64	0	5
06/07/21-155015	06/07/21-155056	41.5226783752441	0.434234946966171	0.81242752075195 3	0.421799868345261	0.820908248424 53	64	0	15
06/07/21-155057	06/07/21-155139	41.7305746078491	0.434160143136978	0.81242752075195 3	0.421479523181915	0.820908248424 53	64	0	20
06/07/21-155139	06/07/21-155221	41.9903509616852	0.434166818857193	0.81293594837188 7	0.422245532274246	0.820908248424 53	64	0	25
06/07/21-155221	06/07/21-155346	84.428061246872	0.434535235166 55	0.81264996528625 5	0.421406567096 71	0.820908248424 53	64	0	30
06/07/21-155556	06/07/21-155721	84.5506846904755	0.437923461198807	0.81199854612350 4	0.422788709402084	0.820908248424 53	64	0.1	20
06/07/21-155721	06/07/21-155846	84.5351805686951	0.436227321624756	0.81309479475021 4	0.421818047761917	0.820908248424 53	64	0.1	25
06/07/21-155846	06/07/21-160011	84.6098339557648	0.435803413391113	0.81260228157043 5	0.421803176403046	0.820908248424 53	64	0.1	30
06/07/21-160011	06/07/21-160053	42.132194519043	0.633819997310638	0.74647688865661 6	0.446623355150223	0.820908248424 53	64	0.5	5
06/07/21-160054	06/07/21-160137	43.5802342891693	0.498594671487808	0.79363214969635	0.432296603918076	0.820908248424 53	64	0.5	10

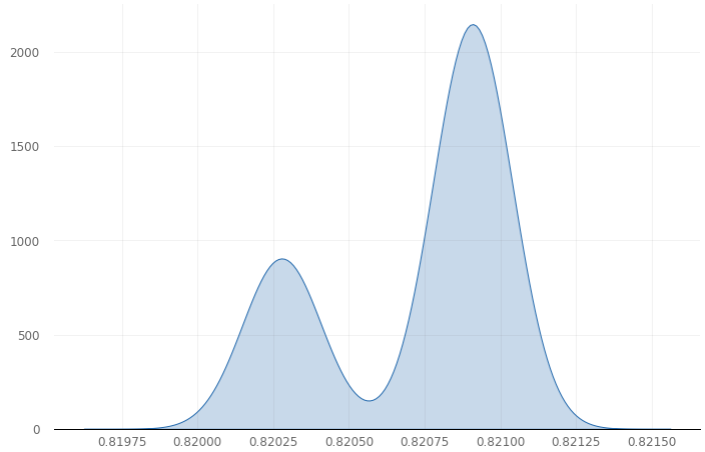
Για καλύτερη οπτικοποίηση των αποτελεσμάτων των πειραμάτων για τις υπερ παραμέτρους παράχθηκαν διάφορες γραφικές όπως και προηγουμένως οι οποίες παρατίθενται πιο κάτω:



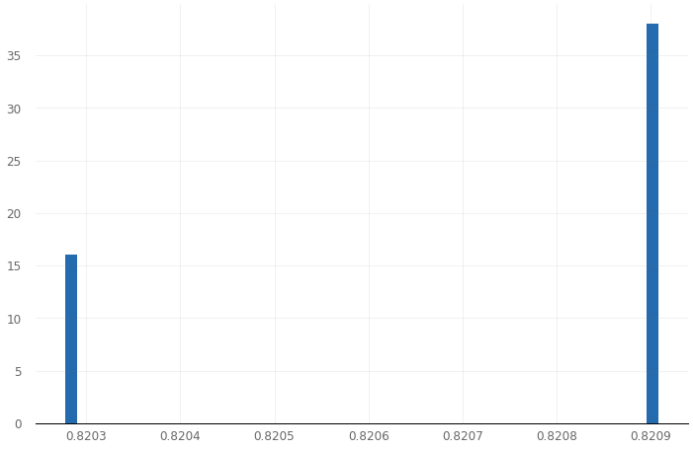
Εικόνα 50 Validation Accuracy vs Validation Loss



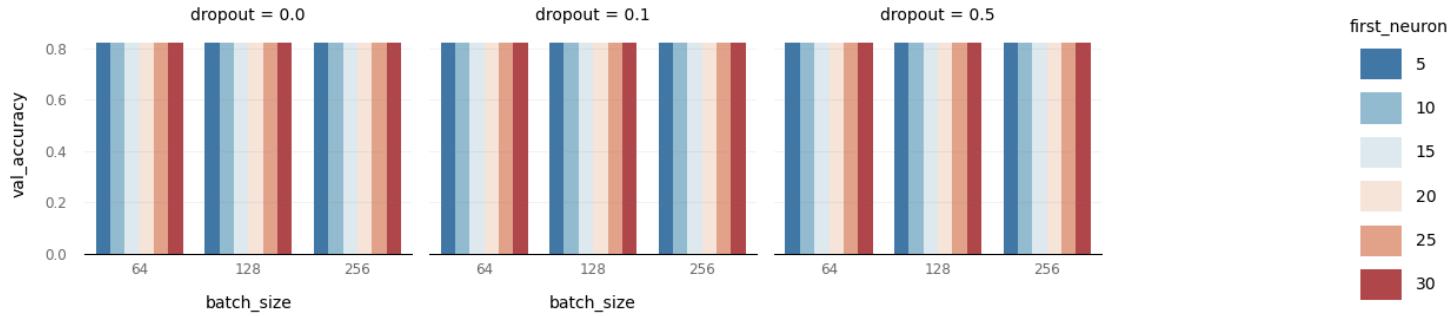
Εικόνα 51 Validation Accuracy of all the experiments



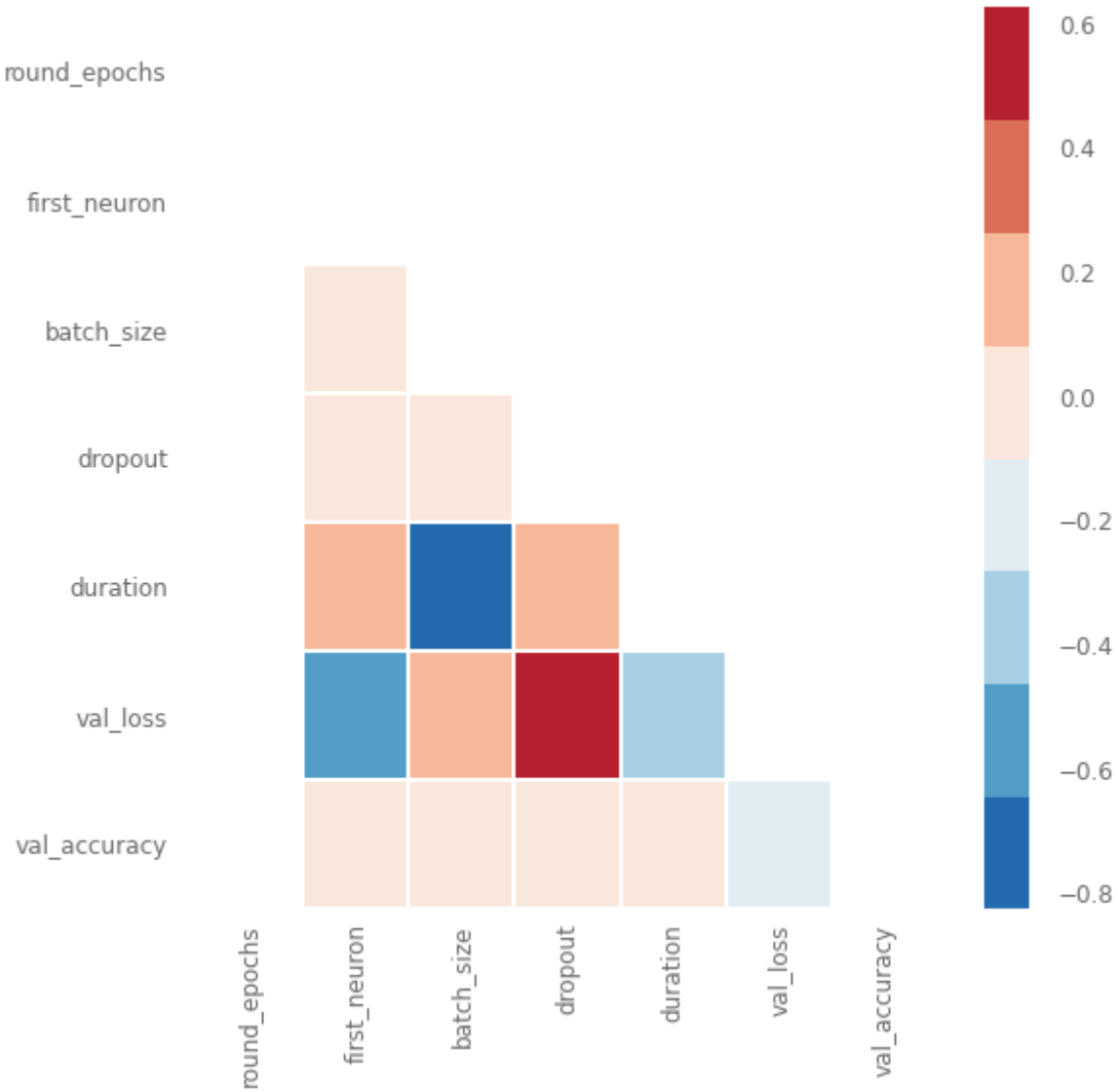
Εικόνα 52 Validation Accuracy of all the experiments - KDE Plot



Εικόνα 53 Ιστόγραμμα Validation Accuracy



Εικόνα 54 Bar Grind



Εικόνα 55 Heatmap Correlation

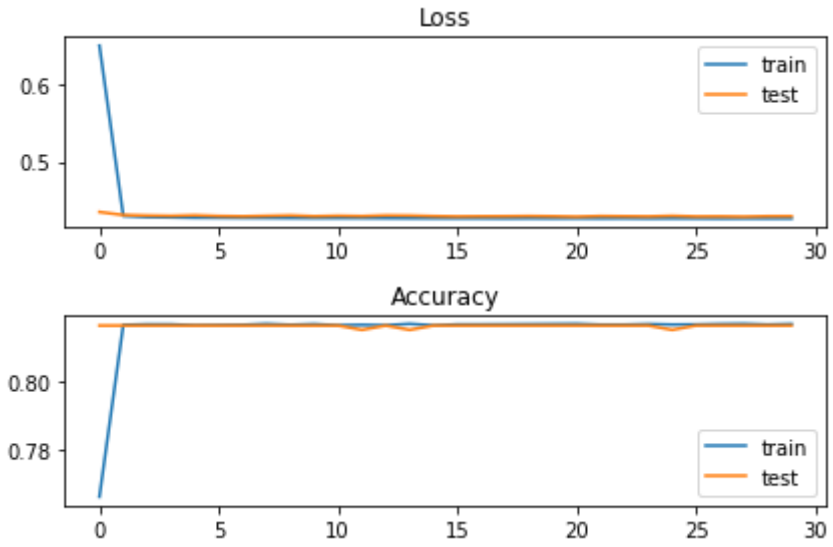
Για τα δεδομένα της προσφοράς δανείου επιλέχθηκαν τελικά οι πιο κάτω υπερ παράμετροι (Πίνακας 14):

Πίνακας 14 Τελική επιλογή υπερ παραμέτρων-2 δραστηριότητες ως είσοδο και 1 ως έξοδο - Offer Dataset

Αριθμός Νευρώνων (neurons)	**20**
Ποσοστό του περιορισμού ενεργοποίησης (dropout)	**0**
Batch Size	**64**

Οι πιο πάνω υπερ παράμετροι επιλέχθηκαν επειδή είχαν εξαιρετικά accuracy και loss και ταυτόχρονα χαμηλό χρόνο εκπαίδευσης των δεδομένων.

Στα επόμενα γραφήματα (Εικόνα 56) φαίνεται η εξέλιξη των loss και accuracy για τις επιλεγμένες υπερ παραμέτρους κατα τη πάροδο της εκπαίδευσης 30 εποχών. Το πρώτο γράφημα δείχνει την απώλεια εγκάρσιας εντροπίας για κάθε εποχή για το σύνολο δεδομένων της εκπαίδευσης (μπλε) και της δοκιμής (πορτοκαλί), και η κάτω γραφική παράσταση δείχνει την ακρίβεια (accuracy) έναντι των εποχών.



Εικόνα 56 Απώλεια εγκάρσιας εντρόπιας και ακρίβεια έναντι των εποχών

5.5.4 Πειράματα για 3 δραστηριότητες ως είσοδο και 1 ως έξοδο - Offer Dataset

Στον ακόλουθο πίνακα (Πίνακας 15) φαίνεται πως όρισα όλες τις παραμέτρους που ήθελα να δοκιμάσει το μοντέλο μου.

Πίνακας 15 Παραμέτροι που θα δοκιμάσει το μοντέλο- 3 δραστηριότητες ως είσοδο και 1 ως έξοδο - Offer Dataset

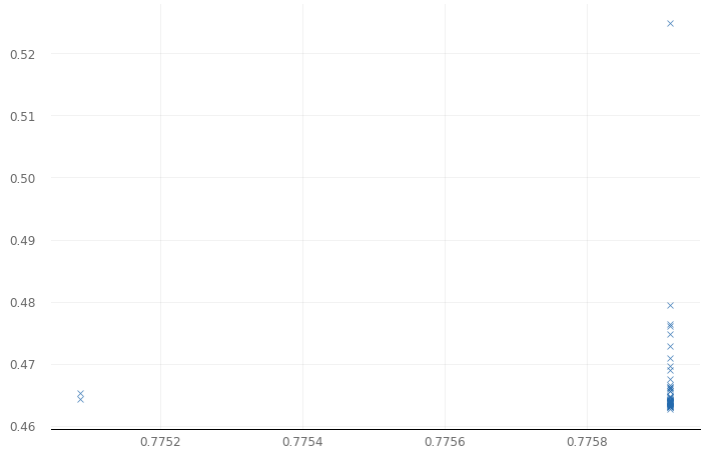
Αριθμός Νευρώνων (neurons)	**[5,10,15,20,25,30]**
Ποσοστό του περιορισμού ενεργοποίησης (dropout)	**[0,0.1,0.5]**
Batch Size	**[32, 64, 128, 256]**

Στον επόμενο πίνακα (Πίνακας 16) παρουσιάζεται ένα δείγμα από τα πρώτα 10 αποτελέσματα των συνολικά 54 πειραμάτων που έγιναν για 15 εποχές στις οποίες παρατηρήθηκε ότι ήταν αρκετές για επίτευξη βέλτιστων αποτελεσμάτων. Ο πίνακας είναι ταξινομημένος σε αύξουσα σειρά με βάση το validation loss.

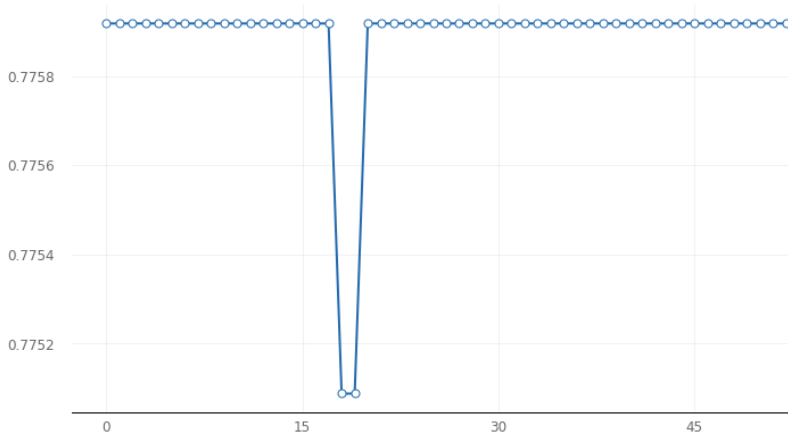
Πίνακας 16 Αποτελέσματα πειραμάτων- 3 δραστηριότητες ως είσοδο και 1 ως έξοδο - Offer Dataset

Start	End	Duration	Loss	Accuracy	val loss	val accuracy	batch size	dropout	first neuron
06/08/21-094844	06/08/21-094922	38.045705318450 9	0.468228489160538	0.773398816585541	0.462662994861603	0.775917470 45517	64	0	20
06/08/21-100120	06/08/21-100144	24.215568780899	0.468725711107254	0.772776782512665	0.463018029928207	0.775917470 45517	128	0	30
06/08/21-094715	06/08/21-094759	43.585777759552	0.468149662017822	0.773309946060181	0.463022887706757	0.775917470 45517	64	0	10
06/08/21-095333	06/08/21-095417	44.425686836 2427	0.470329403877258	0.772443532943726	0.463111072778702	0.775917470 45517	64	0.1	25
06/08/21-100013	06/08/21-100056	43.627913236618	0.468579173088074	0.773643195 62912	0.463144183158875	0.775917470 45517	128	0	20
06/08/21-095125	06/08/21-095204	38.138231039 0472	0.473378509283066	0.772821187973022	0.463198691606522	0.775917470 45517	64	0.1	10
06/08/21-095825	06/08/21-095908	43.575929403 3051	0.476197332143784	0.772910058498383	0.463239043943 9 51035	0.775917470 45517	64	0.5	30
06/08/21-095204	06/08/21-095248	43.554141044 6167	0.472488492727 28	0.772887825965881	0.463412255048752	0.775917470 45517	64	0.1	15
06/08/21-100249	06/08/21-100312	22.667532682 4188	0.472095251083374	0.772510170936585	0.463429063558579	0.775917470 45517	128	0.1	20
06/08/21-095418	06/08/21-095458	40.434428691864	0.469937056303024	0.773398816585541	0.463483124997139	0.775917470 45517	64	0.1	30

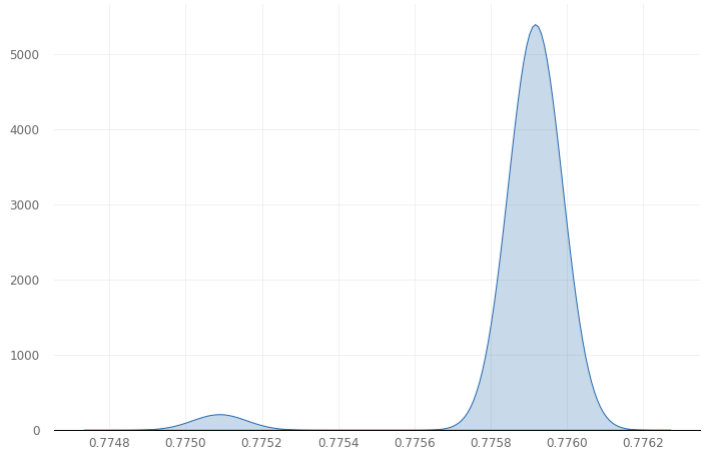
Για καλύτερη οπτικοποίηση των αποτελεσμάτων των πειραμάτων για τις υπερ παραμέτρους παράχθηκαν διάφορες γραφικές όπως και προηγουμένως οι οποίες παρατίθενται πιο κάτω:



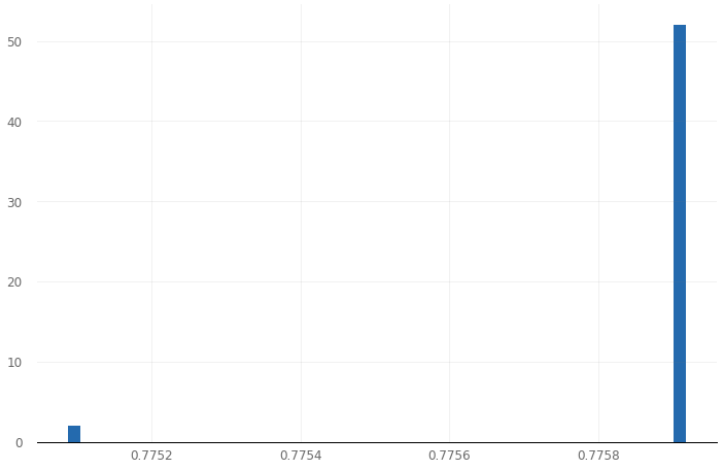
Εικόνα 57 Validation Accuracy vs Validation Loss



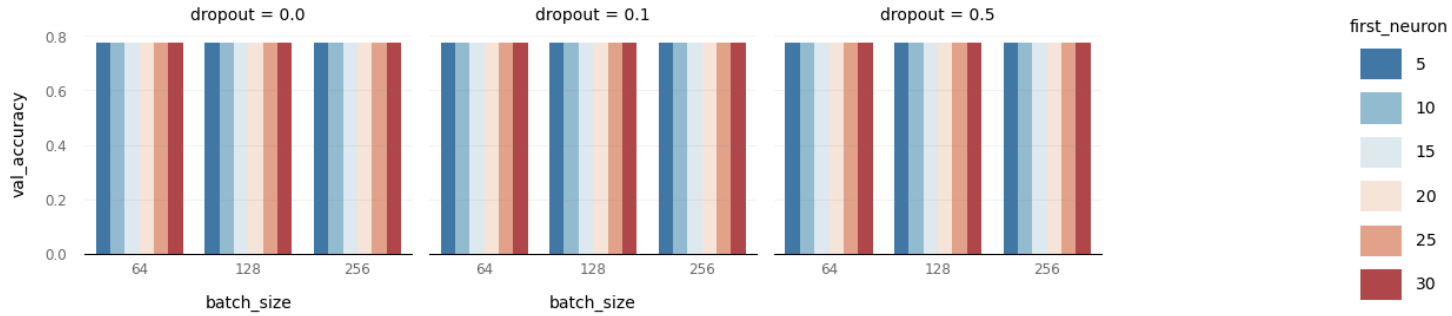
Εικόνα 58 Validation Accuracy of all the experiments



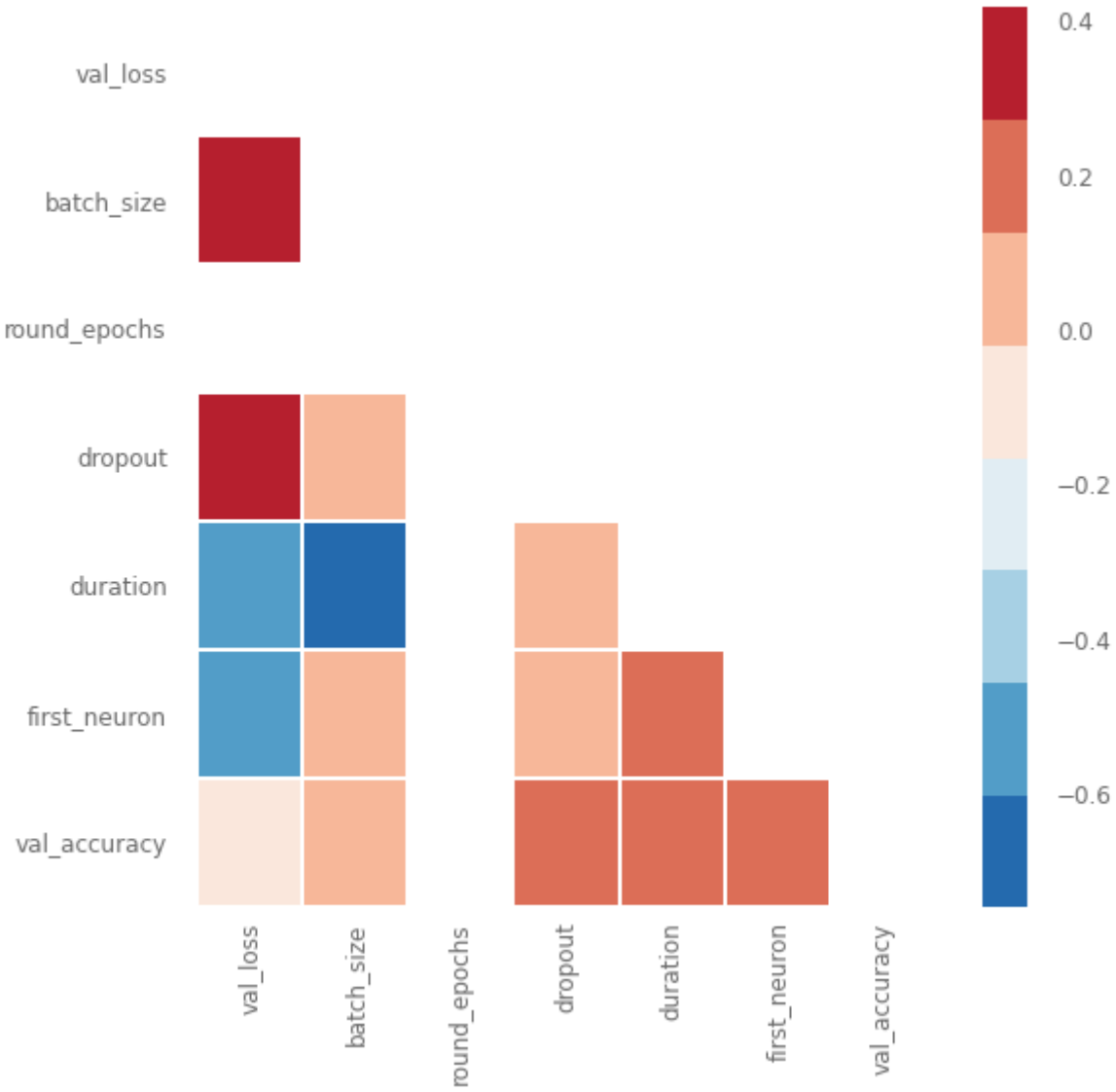
Εικόνα 59 Validation Accuracy of all the experiments - KDE Plot



Εικόνα 60 Ιστόγραμμα Validation Accuracy



Εικόνα 61 Bar Grind



Εικόνα 62 Heatmap Correlation

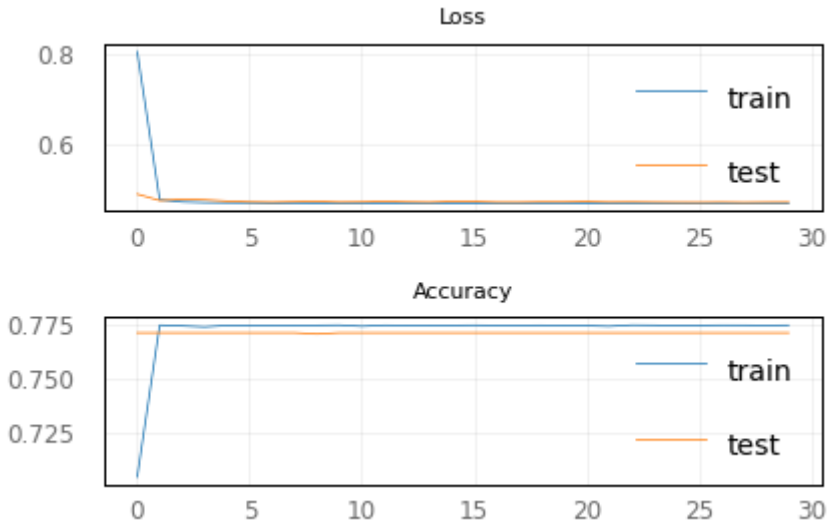
Για τα δεδομένα της προσφοράς δανείου με 3 δραστηριότητες ως είσοδο και 1 ως έξοδο-πρόβλεψη επιλέχθηκαν τελικά οι πιο κάτω υπερ παράμετροι (Πίνακας 17):

Πίνακας 17 Τελική επιλογή υπερ παραμέτρων-3 δραστηριότητες ως είσοδο και 1 ως έξοδο - Offer Dataset

Αριθμός Νευρώνων (neurons)	**30**
Ποσοστό του περιορισμού ενεργοποίησης (dropout)	**0**
Batch Size	**128**

Οι πιο πάνω υπερ παράμετροι επιλέχθηκαν επειδή είχαν εξαιρετικά accuracy και loss και ταυτόχρονα χαμηλό χρόνο εκπαίδευσης των δεδομένων.

Στα επόμενα γραφήματα (Εικόνα 63) φαίνεται η εξέλιξη των loss και accuracy για τις επιλεγμένες υπερ παραμέτρους κατα τη πάροδο της εκπαίδευσης 30 εποχών. Το πρώτο γράφημα δείχνει την απώλεια εγκάρσιας εντροπίας για κάθε εποχή για το σύνολο δεδομένων της εκπαίδευσης (μπλε) και της δοκιμής (πορτοκαλί), και η κάτω γραφική παράσταση δείχνει την ακρίβεια (accuracy) έναντι των εποχών.



Εικόνα 63 Απώλεια εγκάρσιας εντρόπιας και ακρίβεια έναντι των εποχών

Παρατηρούμε ότι, τα γραφήματα δείχνουν ότι το μοντέλο φαίνεται να συγκλίνει. Τα γραφήματα τόσο για την απώλεια εγκάρσιας εντροπίας όσο και για ακρίβεια δείχνουν καλή συμπεριφορά σύγκλισης η οποία ταιριάζει με τα θεωρητικά επιθυμητά αποτελέσματα. Το μοντέλο μπορεί να είναι καλά διαμορφωμένο δεδομένου ότι δεν υπάρχει ένδειξη για underfiiting.

5.6 Αποτελέσματα προβλέψεων

5.6.1 Αποτελέσματα Προβλέψεων (Application Dataset) -2 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Πιο κάτω (Πίνακας 18) παρουσιάζονται τα αποτελέσματα των προβλέψεων όταν δίνουμε 2 δραστηριότητες ως είσοδο στο νευρωνικό και περιμένουμε την αμέσως επόμενη ως έξοδο. Η

μετρική accuracy δείχνει την ακρίβεια των προβλέψεων δηλαδή τις σωστές προβλέψεις διά τις συνολικές προβλέψεις (σωστές και λάθος). Ακόμα στον πίνακα παρουσιάζεται και η ακρίβεια των προβλέψεων εάν λάβουμε υπόψη και τη δεύτερη μεγαλύτερη πιθανότητα που προβλέπει το νευρωνικό. Εάν δηλαδή η δεύτερη πιο μεγάλη πιθανότητα είναι και αναμενόμενη έξοδος τότε βάζουμε το συγκεκριμένο sequence δραστηριοτήτων στις σωστές προβλέψεις. Αυτό έγινε για να δούμε εάν οι λάθος προβλέψεις αποκλίνουν πολύ από τις αναμενόμενες. Παρατηρούμε ότι λαμβάνοντας υπόψη τη δεύτερη πιο μεγάλη πιθανότητα η ακρίβεια εκτοξεύεται από 73,82 % στο 90,17%.

Πίνακας 18 Αποτελέσματα Προβλέψεων (Application Dataset) -2 Δραστηριότητες ως είσοδος και 1 ως έξοδος

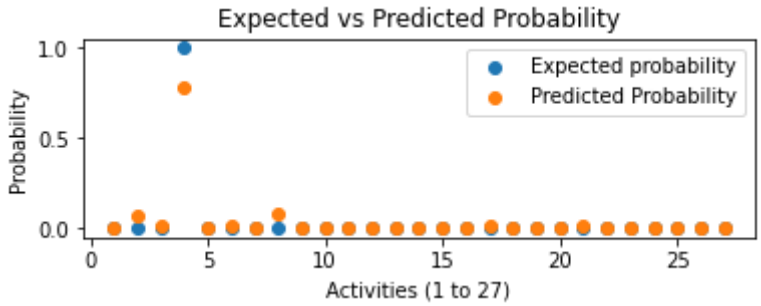
accuracy	**0.73821**
correct predictions	**285126**
false predictions	**101115**
accuracy_secondmax	**0.90170**
correct predictions secondmax	**348274**
false predictions secondmax	**37967**

Στη συνέχεια παρουσιάζονται 3 διαφορετικές περιπτώσεις προβλέψεων του νευρωνικού. Στις δύο λίστες με την αναμενόμενη έξοδο και την πρόβλεψη έχουμε τις 27 δραστηριότητες όπως έχουν οριστεί και προηγουμένως με τους αντίστοιχους ακέραιους αριθμούς.

Στην πρώτη περίπτωση (Πίνακας 19, Εικόνα 64) παρατηρούμε ότι το νευρωνικό προέβλεψε σωστά την 4η δραστηριότητα (w_completeapplication) η οποία ήταν και η αναμενόμενη.

Πίνακας 19 Παράδειγμα σωστής πρόβλεψης (Application Dataset) -2 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Αναμενόμενη Έξοδος	[0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
Πρόβλεψη νευρωνικού	[1.09586341e-03 7.38454312e-02 1.89860910e-02 7.87029803e-01 2.27319179e-05 1.06170066e-02 9.52830276e-14 8.23690593e-02 1.76231072e-06 5.18553245e-13 4.70738041e-06 1.43526404e-05 2.94982223e-04 5.63048388e-06 2.40952033e-03 1.46452294e-05 1.05273742e-02 6.47384219e-13 6.37630874e-05 4.06273730e-11 1.22212302e-02 3.16242819e-07 3.24444729e-04 1.90052215e-05 1.21335266e-04 1.09717412e-05 9.56044044e-09]

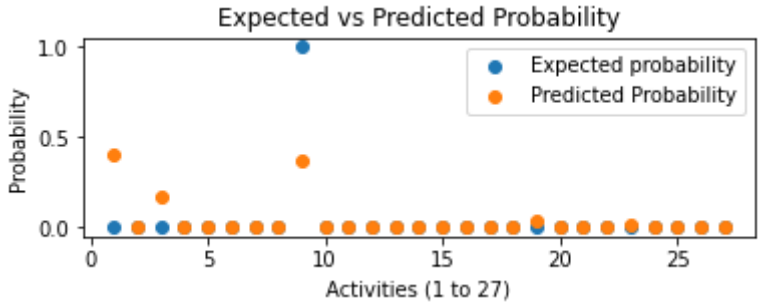


Εικόνα 64 Παράδειγμα σωστής πρόβλεψης (Application Dataset) -2 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Σε αυτή την περίπτωση (Πίνακας 20, Εικόνα 65) παρατηρούμε πως η αναμενόμενη δραστηριότητα ήταν η 9η (a_validating) αλλά το νευρωνικό προέβλεψε ως πιο πιθανή έξοδο την δραστηριότητα 1 (w_validateapplication). Εδώ όμως συμβαίνει αυτό που αναφέρθηκε και πιο πριν: η δεύτερη πιο μεγάλη πιθανότητα που προέβλεψε το νευρωνικό ήταν και η σωστή . Η πιθανότητα αυτή ήταν μάλιστα και πολύ κοντά στην πιθανότητα της δραστηριότητας 1.

Πίνακας 20 Παράδειγμα σωστής πρόβλεψης με τη δεύτερη πιο μεγάλη πιθανότητα (Application Dataset) -2 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Αναμενόμενη Έξοδος	[0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
Πρόβλεψη νευρωνικού	[3.99111867e-01 3.45084743e-07 1.71036795e-01 2.09409918e-05 5.70161228e-06 2.30285674e-04 9.69248015e-10 6.56255861e-06 3.68090451e-01 5.94366001e-10 3.60389231e-06 2.55426421e-05 3.78487130e-05 2.02667252e-05 3.57575668e-03 1.78497266e-05 4.98886475e-05 5.23009691e-10 3.88361886e-02 6.42769615e-10 1.82062329e-04 2.58714635e-06 1.61367562e-02 2.59229587e-03 5.69897020e-06 3.83759470e-07 1.02863305e-05]

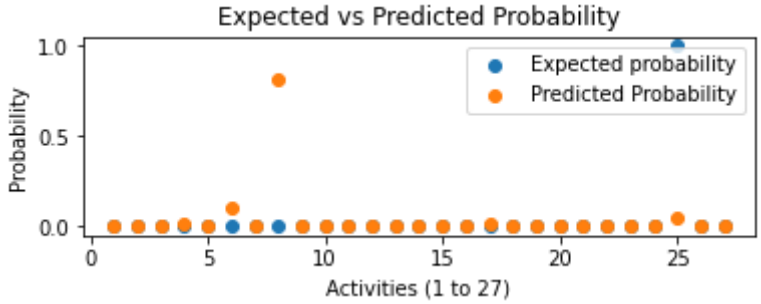


Εικόνα 65 Παράδειγμα σωστής πρόβλεψης με τη δεύτερη πιο μεγάλη πιθανότητα (Application Dataset) -2 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Στην τρίτη περίπτωση (Πίνακας 21, Εικόνα 66) παρατηρούμε πως η αναμενόμενη έξοδος ήταν η δραστηριότητα 25 (o_sent(onlineonly)). Το νευρωνικό δεν κατάφερε όμως να την προβλέψει ούτε ως πρώτη πιο μεγάλη πιθανότητα ούτε ως δεύτερη. Ήταν όμως με μικρή διαφορά από την δεύτερη πιο μεγάλη η τρίτη πιο πιθανή πρόβλεψη.

Πίνακας 21 Παράδειγμα λάθος πρόβλεψης (Application Dataset) -2 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Αναμενόμενη Έξοδος	[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0]
Πρόβλεψη νευρωνικού	[8.3310486e-05 9.0894266e-04 5.4848415e-04 1.4773463e-02 1.4199615e-05 9.8382212e-02 8.6540499e-09 8.1331110e-01 1.4045164e-05 1.8377838e-09 3.0430947e-05 6.7525078e-05 4.6772766e-05 5.0415010e-06 5.5322842e-05 1.1655137e-05 1.6707877e-02 1.9876905e-09 2.2037178e-05 1.6880769e-09 1.3242485e-03 1.8634593e-05 4.2294373e-04 1.0113612e-05 5.3189233e-02 5.2120402e-05 3.0356969e-07]



Εικόνα 66 Παράδειγμα λάθος πρόβλεψης (Application Dataset) -2 Δραστηριότητες ως είσοδος και 1 ως έξοδος

5.6.2 Αποτελέσματα Προβλέψεων (Application Dataset) -3 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Πιο κάτω (Πίνακας 22) παρουσιάζονται τα αποτελέσματα των προβλέψεων όταν δίνουμε 3 δραστηριότητες ως είσοδο στο νευρωνικό και περιμένουμε την αμέσως επόμενη ως έξοδο. Βλέπουμε πως σε αυτή την περίπτωση η ακρίβεια είναι πολύ υψηλότερη σε σχέση με την προηγούμενη. Επίσης, παρατηρούμε ότι λαμβάνοντας υπόψη τη δεύτερη πιο μεγάλη πιθανότητα η ακρίβεια εκτοξεύεται στο 96,01% .

Πίνακας 22 Αποτελέσματα Προβλέψεων (Application Dataset) -3 Δραστηριότητες ως είσοδος και 1 ως έξοδος

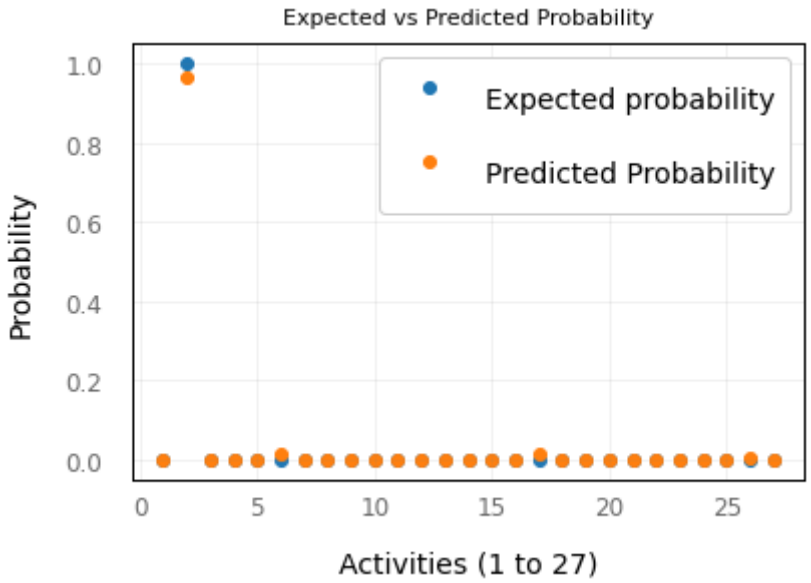
accuracy	**0.83374**
correct predictions	**313356**
false predictions	**62490**
accuracy_secondmax	**0.96012**
correct predictions secondmax	**360858**
false predictions secondmax	**14988**

Στη συνέχεια παρουσιάζονται 3 διαφορετικές περιπτώσεις προβλέψεων του νευρωνικού. Στις δύο λίστες με την αναμενόμενη έξοδο και την πρόβλεψη έχουμε τις 27 δραστηριότητες όπως έχουν οριστεί και προηγουμένως με τους αντίστοιχους ακέραιους αριθμούς.

Στην πρώτη περίπτωση (Πίνακας 23, Εικόνα 67)παρατηρούμε ότι το νευρωνικό προέβλεψε σωστά την 2η δραστηριότητα (w_callafteroffers) η οποία ήταν και η αναμενόμενη.

Πίνακας 23 Παράδειγμα σωστής πρόβλεψης (Application Dataset) 3 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Αναμενόμενη Έξοδος	[0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
Πρόβλεψη νευρωνικού	1.07927546e-04 9.64808345e-01 9.81179141e-07 5.89762249e-06 2.65384354e-08 1.51019087e-02 2.30492976e-08 1.41370064e-03 1.35970228e-12 1.37641454e-09 1.07731694e-05 3.42400135e-08 3.02376811e-05 1.20190606e-07 6.47144816e-06 1.46401086e-11 1.27453478e-02 8.83036644e-10 3.01243318e-07 2.15753578e-11 1.97946350e-03 9.67799042e-06 5.32371283e-04 2.26718930e-05 5.57262774e-06 3.21817235e-03 5.12302867e-10

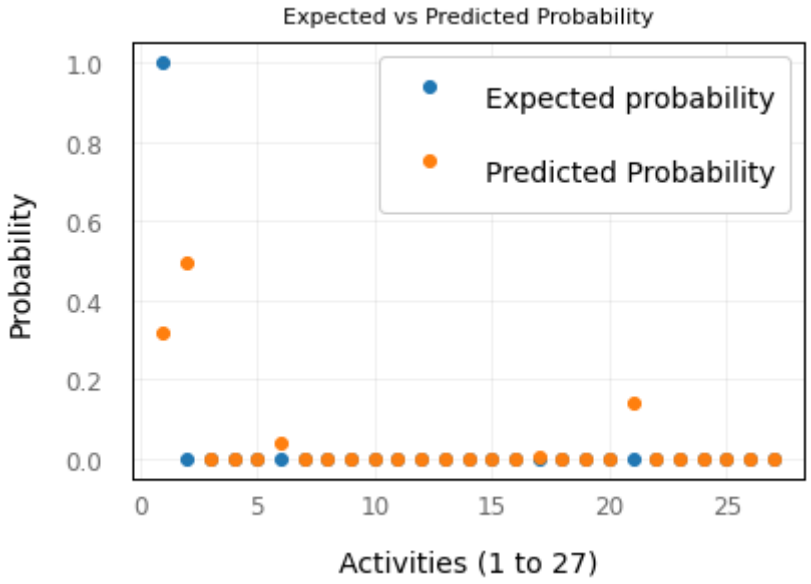


Εικόνα 67 Παράδειγμα σωστής πρόβλεψης (Application Dataset) -3 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Σε αυτή την περίπτωση (Πίνακας 24, Εικόνα 68) παρατηρούμε πως η αναμενόμενη δραστηριότητα ήταν η 1η (w_validateapplication) αλλά το νευρωνικό προέβλεψε ως πιο πιθανή έξοδο την δραστηριότητα 2 (w_callafteroffers). Εδώ όμως συμβαίνει αυτό που αναφέρθηκε και πιο πριν: η δεύτερη πιο μεγάλη πιθανότητα που προέβλεψε το νευρωνικό ήταν και η σωστή . Η πιθανότητα αυτή ήταν μάλιστα και πολύ κοντά στην πιθανότητα της δραστηριότητας 1.

Πίνακας 24 Παράδειγμα σωστής πρόβλεψης με τη δεύτερη πιο μεγάλη πιθανότητα (Application Dataset) -3 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Αναμενόμενη Έξοδος	[1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
Πρόβλεψη νευρωνικού	[3.2025120e-01 4.9361426e-01 1.0841135e-05 7.9621159e-06 5.1000154e-07 4.1347571e-02 1.1384365e-12 6.7785790e-05 1.6839521e-11 3.7994437e-11 5.6435956e-06 5.0256995e-07 8.4329053e-04 1.5001532e-08 5.5780511e-05 2.0031077e-09 2.3482265e-03 7.0899057e-11 3.6360678e-05 7.0123938e-09 1.4015906e-01 1.9813906e-06 8.8808965e-04 3.2783163e-04 5.5737695e-07 3.2498989e-05 5.6343347e-10]

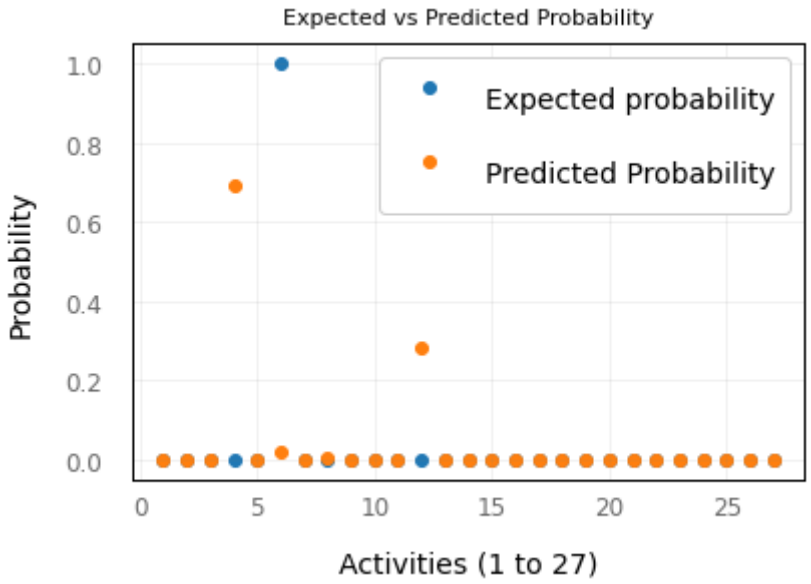


Εικόνα 68 Παράδειγμα σωστής πρόβλεψης με τη δεύτερη πιο μεγάλη πιθανότητα (Application Dataset) -3 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Στην τρίτη περίπτωση (Πίνακας 25, Εικόνα 69)παρατηρούμε πως η αναμενόμενη έξοδος ήταν η δραστηριότητα 6 (o_createoffer). Το νευρωνικό δεν κατάφερε όμως να την προβλέψει ούτε ως πρώτη πιο μεγάλη πιθανότητα ούτε ως δεύτερη. Ήταν όμως με μικρή διαφορά από την δεύτερη πιο μεγάλη η τρίτη πιο πιθανή πρόβλεψη.

Πίνακας 25 Παράδειγμα λάθος πρόβλεψης (Application Dataset) -3 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Αναμενόμενη Έξοδος	[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
Πρόβλεψη νευρωνικού	[1.0318197e-05 7.5462694e-06 8.2994402e-06 6.9226319e-01 1.4791372e-07 1.8551346e-02 8.7929159e-08 3.5538338e-03 4.1091588e-12 1.3894705e-09 1.6817423e-05 2.8421712e-01 2.6148815e-08 2.0903985e-09 6.0550115e-06 1.1636716e-11 2.5841649e-04 1.1977161e-09 3.3524157e-05 8.8982475e-09 5.4777646e-04 3.7993839e-06 1.8429459e-04 1.9573487e-04 7.8283760e-05 6.3303320e-05 2.8686172e-08]



Εικόνα 69 Παράδειγμα λάθος (Application Dataset) -3 Δραστηριότητες ως είσοδος και 1 ως έξοδος

5.6.3 Αποτελέσματα Προβλέψεων (Offer Dataset) -2 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Πιο κάτω (Πίνακας 26) παρουσιάζονται τα αποτελέσματα των προβλέψεων όταν δίνουμε 2 δραστηριότητες ως είσοδο στο νευρωνικό και περιμένουμε την αμέσως επόμενη ως έξοδο. Παρατηρούμε ότι λαμβάνοντας υπόψη τη δεύτερη πιο μεγάλη πιθανότητα η ακρίβεια εκτοξεύεται στο 96,86% από 81,60% που είναι υπό κανονικές συνθήκες.

Πίνακας 26 Αποτελέσματα Προβλέψεων (Offer Dataset) -2 Δραστηριότητες ως είσοδος και 1 ως έξοδος

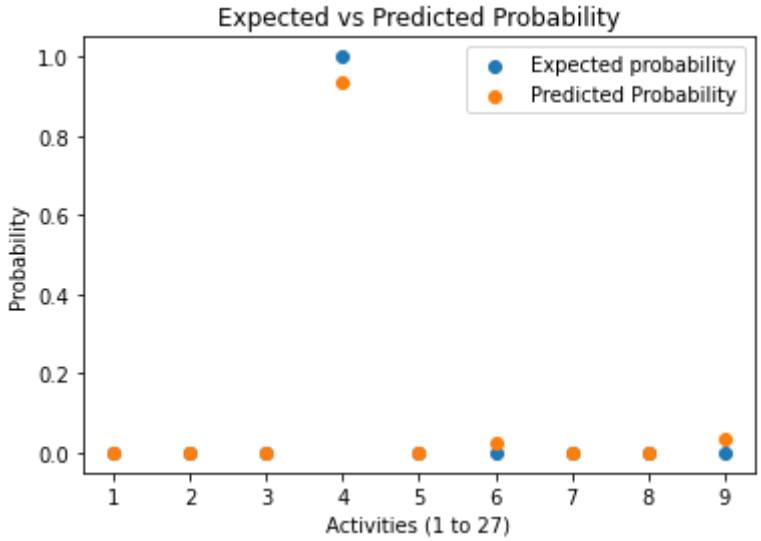
accuracy	**0.8160310591660938**
correct predictions	**40356**
false predictions	**9098**
accuracy_secondmax	**0.9686173009261132**
correct predictions secondmax	**47902**
false predictions secondmax	**1552**

Στη συνέχεια παρουσιάζονται 3 διαφορετικές περιπτώσεις προβλέψεων του νευρωνικού. Στις δύο λίστες με την αναμενόμενη έξοδο και την πρόβλεψη έχουμε τις 9 δραστηριότητες όπως έχουν οριστεί και προηγουμένως με τους αντίστοιχους ακέραιους αριθμούς.

Στην πρώτη περίπτωση (Πίνακας 27, Εικόνα 70) παρατηρούμε ότι το νευρωνικό προέβλεψε σωστά την 4η δραστηριότητα (o_sent(mailandonline)) η οποία ήταν και η αναμενόμενη.

Πίνακας 27 Παράδειγμα σωστής πρόβλεψης (Offer Dataset) -2 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Αναμενόμενη Έξοδος	[0 0 0 1 0 0 0 0 0]
Πρόβλεψη νευρωνικού	[1.0374034e-09 1.0437367e-09 9.4605412e-06 9.3486696e-01 2.0870334e-06 2.7054617e-02 4.5771820e-05 1.2595387e-03 3.6761642e-02]



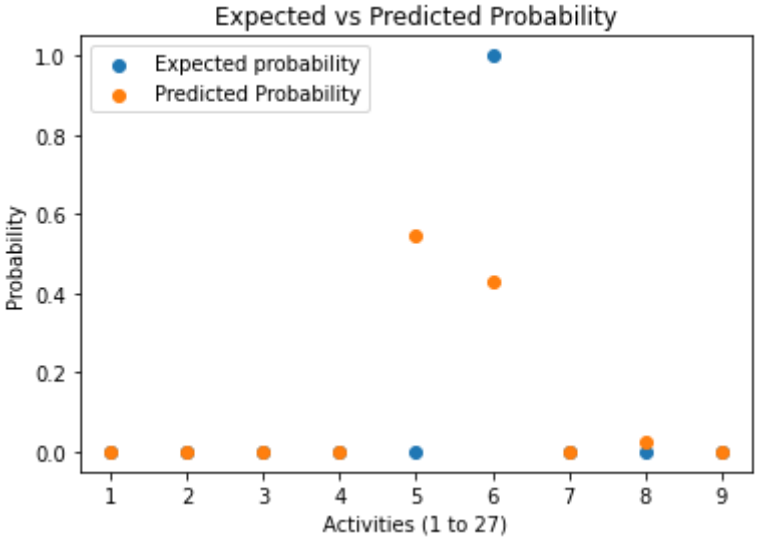
Εικόνα 70 Παράδειγμα σωστής πρόβλεψης (Offer Dataset) -2 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Σε αυτή την περίπτωση (Πίνακας 28, Εικόνα 71) παρατηρούμε πως η αναμενόμενη δραστηριότητα ήταν η 6η (o_cancelled) αλλά το νευρωνικό προέβλεψε ως πιο πιθανή έξοδο

την δραστηριότητα 7η (o_accepted). Εδώ όμως συμβαίνει αυτό που αναφέρθηκε και πιο πριν: η δεύτερη πιο μεγάλη πιθανότητα που προέβλεψε το νευρωνικό ήταν και η σωστή . Η πιθανότητα αυτή ήταν μάλιστα και πολύ κοντά στην πιθανότητα της δραστηριότητας 1.

Πίνακας 28 Παράδειγμα σωστής πρόβλεψης με τη δεύτερη πιο μεγάλη πιθανότητα (Offer Dataset) -2 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Αναμενόμενη Έξοδος	[0 0 0 0 0 1 0 0 0]
Πρόβλεψη νευρωνικού	[8.5081459e-10 8.5469870e-10 9.9437276e-04 9.6083086e-06 5.4722774e-01 4.2816332e-01 6.1562387e-05 2.3536410e-02 6.9658104e-06]

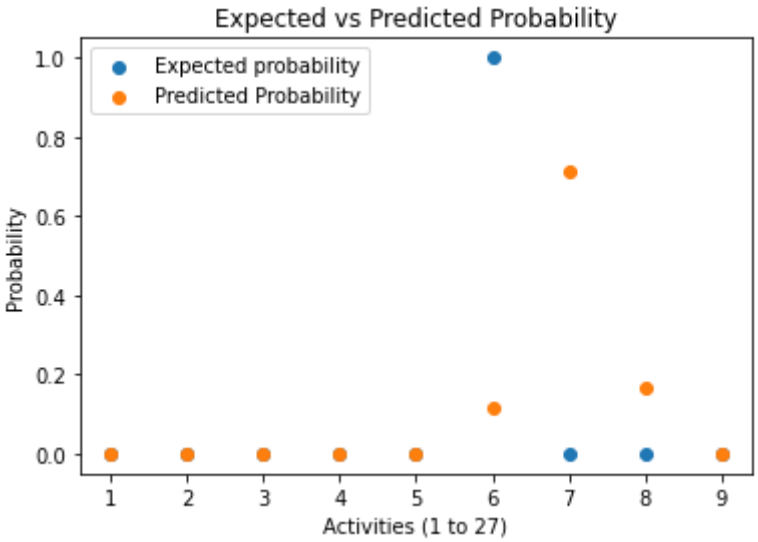


Εικόνα 71 Παράδειγμα σωστής πρόβλεψης με τη δεύτερη πιο μεγάλη πιθανότητα (Offer Dataset) -2 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Στην τρίτη περίπτωση (Πίνακας 29, Εικόνα 72) παρατηρούμε πως η αναμενόμενη έξοδος ήταν η δραστηριότητα 6 (o_cancelled). Το νευρωνικό δεν κατάφερε όμως να την προβλέψει ούτε ως πρώτη πιο μεγάλη πιθανότητα ούτε ως δεύτερη. Ήταν όμως με μικρή διαφορά από την δεύτερη πιο μεγάλη η τρίτη πιο πιθανή πρόβλεψη.

Πίνακας 29 Παράδειγμα λάθος πρόβλεψης (Offer Dataset) -2 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Αναμενόμενη Έξοδος	[0 0 0 0 0 1 0 0 0]
Πρόβλεψη νευρωνικού	[1.1897001e-09 1.0481239e-09 1.0806430e-03 8.3887517e-06 4.0700629e-06 1.1692060e-01 7.1434462e-01 1.6763774e-01 3.9714901e-06]



Εικόνα 72 Παράδειγμα λάθος πρόβλεψης (Offer Dataset) -2 Δραστηριότητες ως είσοδος και 1 ως έξοδος

5.6.4 Αποτελέσματα Προβλέψεων (Offer Dataset) -3 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Πιο κάτω παρουσιάζονται τα αποτελέσματα των προβλέψεων όταν δίνουμε 3 δραστηριότητες ως έξοδο στο νευρωνικό και περιμένουμε την αμέσως επόμενη ως έξοδο. Βλέπουμε πως σε αυτή την περίπτωση η ακίβεια έχει μειωθεί σε σχέση με αύτη που είχαμε στις 2 δραστηριότητες ως είσοδο στο offer dataset. Παρατηρούμε επίσης ότι λαμβάνοντας υπόψη τη δεύτερη πιο μεγάλη πιθανότητα η ακρίβεια εκτοξεύεται στο 96,64%.

Πίνακας 30 Αποτελέσματα Προβλέψεων (Offer Dataset) -3 Δραστηριότητες ως είσοδος και 1 ως έξοδος

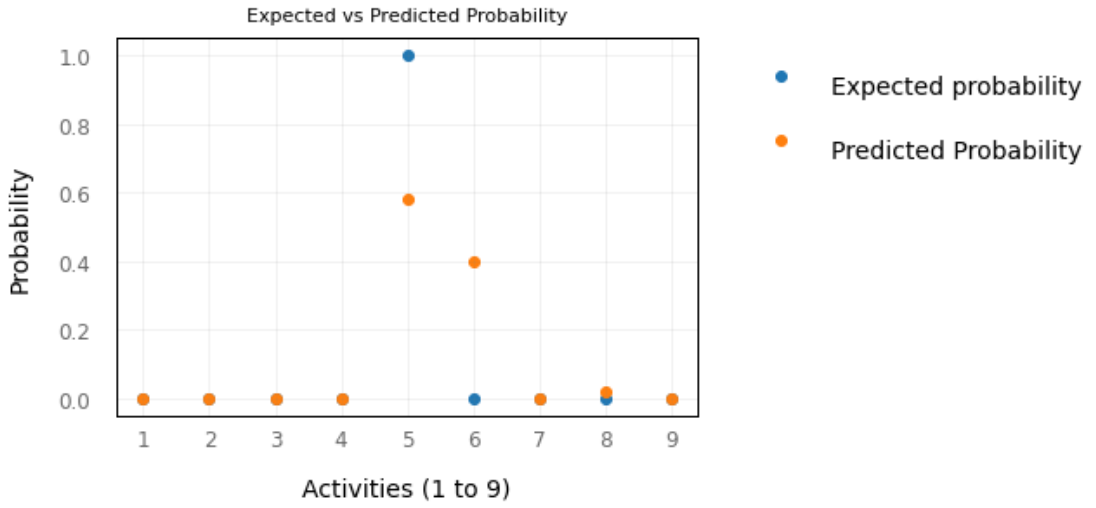
accuracy	0.77078
correct predictions	27261
false predictions	8107
accuracy_secondmax	0.96641
correct predictions secondmax	34180
false predictions secondmax	1188

Στη συνέχεια παρουσιάζονται 3 διαφορετικές περιπτώσεις προβλέψεων του νευρωνικού. Στις δύο λίστες με την αναμενόμενη έξοδο και την πρόβλεψη έχουμε τις 9 δραστηριότητες όπως έχουν οριστεί και προηγουμένως με τους αντίστοιχους ακέραιους αριθμούς.

Στην πρώτη περίπτωση (Πίνακας 31, Εικόνα 73) παρατηρούμε ότι το νευρωνικό προέβλεψε σωστά την 5η δραστηριότητα (o_returned) η οποία ήταν και η αναμενόμενη.

Πίνακας 31 Παράδειγμα σωστής πρόβλεψης (Offer Dataset) -3 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Αναμενόμενη Έξοδος	[0 0 0 0 1 0 0 0 0]
Πρόβλεψη νευρωνικού	[3.2243602e-08 2.5280936e-08 7.1780250e-04 3.2240155e-08 5.7977116e-01 3.9786693e-01 5.3498497e-05 2.1590488e-02 3.0074236e-08]

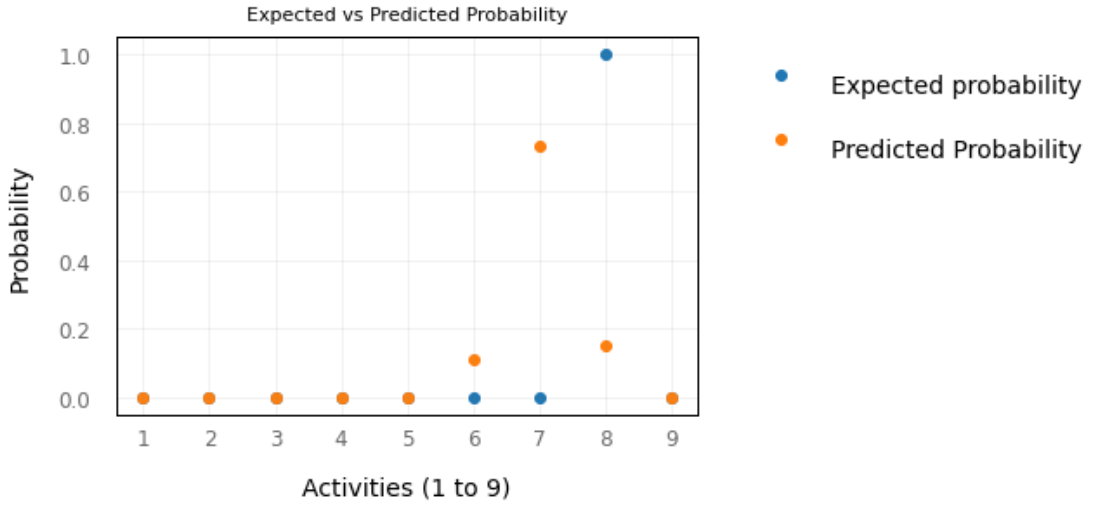


Εικόνα 73 Παράδειγμα σωστής πρόβλεψης (Offer Dataset) -3 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Σε αυτή την περίπτωση (Πίνακας 32, Εικόνα 74) παρατηρούμε πως η αναμενόμενη δραστηριότητα ήταν η 6η (o_cancelled) αλλά το νευρωνικό προέβλεψε ως πιο πιθανή έξοδο την δραστηριότητα 7η (o_accepted). Εδώ όμως συμβαίνει αυτό που αναφέρθηκε και πιο πριν: η δεύτερη πιο μεγάλη πιθανότητα που προέβλεψε το νευρωνικό ήταν και η σωστή .

Πίνακας 32 Παράδειγμα σωστής πρόβλεψης με τη δεύτερη πιο μεγάλη πιθανότητα (Offer Dataset) -3 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Αναμενόμενη Έξοδος	[0 0 0 0 0 0 0 1 0]
Πρόβλεψη νευρωνικού	[5.81352424e-08 3.97982092e-08 1.91099755e-03 4.73188209e-08 2.09382924e-04 1.13231644e-01 7.34072983e-01 1.50574774e-01 5.02159878e-08]

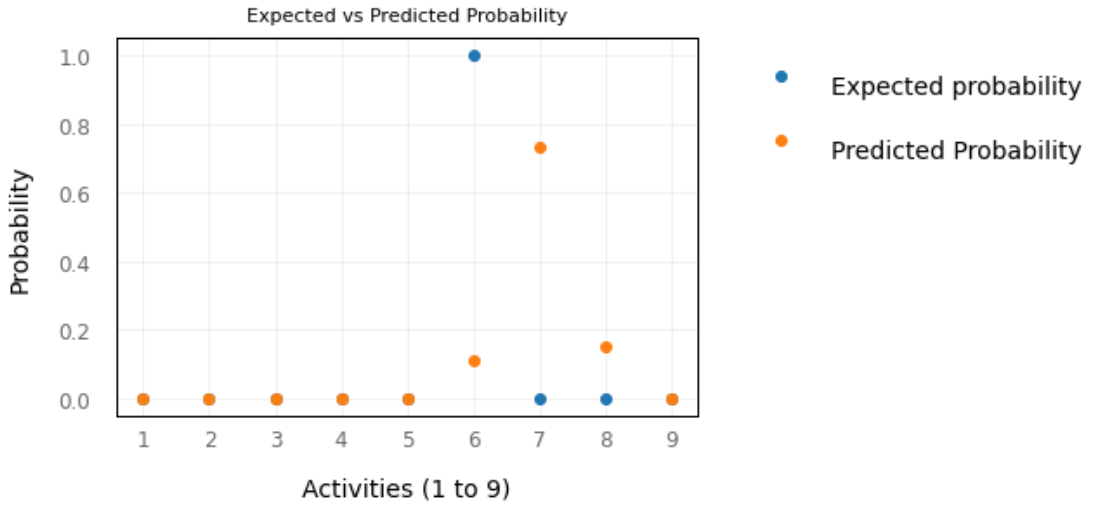


Εικόνα 74 Παράδειγμα σωστής πρόβλεψης με τη δεύτερη πιο μεγάλη πιθανότητα (Offer Dataset) -3 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Στην τρίτη περίπτωση (Πίνακας 33, Εικόνα 75) παρατηρούμε πως η αναμενόμενη έξοδος ήταν η δραστηριότητα 6 (o_cancelled). Το νευρωνικό δεν κατάφερε όμως να την προβλέψει ούτε ως πρώτη πιο μεγάλη πιθανότητα ούτε ως δεύτερη. Ήταν όμως με μικρή διαφορά από την δεύτερη πιο μεγάλη η τρίτη πιο πιθανή πρόβλεψη.

Πίνακας 33 Παράδειγμα λάθος (Offer Dataset) -3 Δραστηριότητες ως είσοδος και 1 ως έξοδος

Αναμενόμενη Έξοδος	[0 0 0 0 0 1 0 0 0]
Πρόβλεψη νευρωνικού	[5.81352424e-08 3.97982092e-08 1.91099755e-03 4.73188209e-08 2.09382924e-04 1.13231644e-01 7.34072983e-01 1.50574774e-01 5.02159878e-08]



Εικόνα 75 Παράδειγμα λάθος (Offer Dataset) -3 Δραστηριότητες ως είσοδος και 1 ως έξοδος

5.7 Συμπεράσματα

5.7.1 Συγκεντρωτικά Αποτελέσματα

Αρχικά από τα πειράματα που έγιναν στις υπερ παραμέτρους παρατηρήθηκε ότι δεν υπήρχαν μεγάλες διαφορές στις τιμές των accuracy και loss. Οπότε ενώ αρχικά η επιλογή των τελικών υπερ παραμέτρων θα βασιζόταν σε αυτά και ενώ σε κάποιες περιπτώσεις υπάρχουν επιλογές με ελάχιστα καλύτερο συνδυασμό των δύο μετρικών, δεν προτιμήθηκαν. Αυτό έγινε αφού το χρονικό διάστημα που χρειάζονταν τα δεδομένα για να γίνουν train ήταν πολύ υψηλότερο (διπλάσιο και τριπλάσιο ορισμένες φορές) σε σχέση με άλλους συνδυασμούς υπερ παραμέτρων οι οποίοι είχαν μερικά δεκαδικά διαφορά στις μετρικές αλλά ήταν πολύ πιο γρήγοροι κατά τη διάρκεια της εκπαίδευσης. Έτσι, τελικά προτιμήθηκαν σε όλα τα πειράματα οι συνδυασμοί οι οποίοι έχουν εξαιρετικά accuracy και loss και ταυτόχρονα χαμηλό χρόνο εκπαίδευσης των δεδομένων.

Στον επόμενο πίνακα (Πίνακας 34) παρουσιάζονται συγκεντρωμένα τα αποτελέσματα για κάθε dataset και διαφορετικό συνδυασμό δραστηριοτήτων ως είσοδο-έξοδο. Για το dataset της αίτησης δανείου (application) παρατηρούμε ότι η ακρίβεια (accuracy) βελτιώνεται αισθητά όταν η είσοδος αποτελείται από 3 δραστηριότητες αντί 2. Συγκεκριμένα από το 73,8 % εκτοξεύεται στο 83,4%. Αυτό πιθανόν να συμβαίνει επειδή το dataset της αίτησης δανείου αποτελείται από πολλές δραστηριότητες (26), είναι πιο περίπλοκο και γι' αυτό τον λόγο η εξάρτηση της προηγούμενης δραστηριότητας με την επόμενη μπορεί να είναι μεγαλύτερη. Έτσι η αλυσίδα 3 δραστηριοτήτων οδηγεί σε καλύτερα αποτελέσματα από την αλυσίδα 2 δραστηριοτήτων. Ακόμα παρατηρούμε ότι η ακρίβεια second max (η ακρίβεια των προβλέψεων εάν λάβουμε υπόψη και τη δεύτερη μεγαλύτερη πιθανότητα που προβλέπει το νευρωνικό), βελτιώνεται αισθητά και από 90,2% στην είσοδο με 2 δραστηριότητες γίνεται 96% στην είσοδο με 3 δραστηριότητες.

Προχωρώντας στα αποτελέσματα του dataset της προσφοράς δανείου (offer), παρατηρούμε ότι συμβαίνει το αντίθετο από αυτό που συνέβαινε στο προηγούμενο dataset. Εδώ ενώ αρχικά η ακρίβεια είναι 81,6% μειώνεται σε 77,1% όταν οι δραστηριότητες εισόδου γίνονται 3. Αυτό μπορεί να συμβαίνει επειδή σε αντίθεση με τα δεδομένα της αίτησης, τα δεδομένα της προσφοράς αποτελούνται από μόνο 8 δραστηριότητες και η εξάρτηση της επόμενης με την προηγούμενη δραστηριότητα ίσως να μην είναι τόσο μεγάλη. Η second max ακρίβεια σε αυτό το dataset παραμένει σχετικά σταθερή και μειώνεται ελάχιστα από 96,9% στις 2 δραστηριότητες ως είσοδο σε 96,64% στις 3.

Dataset	Είσοδος - Έξοδος	Accuracy	Accuracy (Second Max)
Application Dataset	Είσοδος : 2 δραστηριότητες Έξοδος: 1 δραστηριότητα	0.73821	0.90170
	Είσοδος : 3 δραστηριότητες Έξοδος:1 δραστηριότητα	0.83374	0.96012
Offer Dataset	Είσοδος : 2 δραστηριότητες Έξοδος: 1 δραστηριότητα	0.81603	0.96862
	Είσοδος : 3 δραστηριότητες Έξοδος: 1 δραστηριότητα	0.77078	0.96641

Πίνακας 34 Συγκεντρωτικά αποτελέσματα

5.7.2 Σύγκριση με την Βιβλιογραφία

Λόγω του γεγονότος ότι το dataset που χρησιμοποιώ στην διπλωματική αυτή είναι σχετικά καινούριο δεν υπάρχει δημοσιευμένη δουλειά, με παρόμοιες τεχνικές LSTM και ακριβώς ίδια δεδομένα. Υπάρχει όμως εκτεταμένη βιβλιογραφία με την παλαιότερη εκδοχή του dataset από το Business Process Intelligence Challenge του 2012 (BPIC'12) που επίσης περιγράφει τη διαδικασία υποβολής αιτήσεων δανείου σε ένα Ολλανδικό Οικονομικό Ινστιτούτο.

Στον επόμενο πίνακα (Πίνακας 35) φαίνεται η σύγκριση που έκαναν στο (Efrén *et al.*, 2021) ανάμεσα σε διαφορετικές δημοσιεύσεις στον τομέα της πρόβλεψης της επόμενης δραστηριότητας και χρησιμοποιώντας διαφορετικά σύνολα δεδομένων.

	Helpdesk	BPI 2012	BPI 2012 Complete	BPI 2012 W	BPI 2012 W Complete	BPI 2012 O	BPI 2012 A	BPI 2013 closed problems	BPI 2013 Incidents	Sepsis	Env permit	Nasa
Pasquadibisceglie et al.	65.84	82.59	74.55	81.59	66.14	77.51	71.47	24.35	31.10	56.71	**91.47**	87.96
Tax el al.	75.06	**85.20**	**79.39**	84.90	67.80	81.22	77.75	**65.57**	67.50	**65.87**	89.24	88.15
Camargo et al.	**76.51**	83.41	79.22	83.29	65.19	**85.13**	**78.92**	60.62	68.01	-	**91.38**	-
Hinkka el al.	**77.90**	**86.05**	**79.76**	83.52	67.24	**85.51**	**79.27**	**61.14**	**77.95**	64.44	89.46	87.89
Khan et al.	69.13	82.93	75.50	**86.69**	**75.91**	84.48	75.62	55.57	64.34	64.34	89.78	85.51
Evermann et al.	70.07	60.38	63.37	75.22	65.38	79.20	74.44	55.66	68.15	34.37	84.33	20.43
Mauro et al.	74.77	84.56	78.72	85.11	65.01	81.52	78.09	56.97	**71.09**	**64.82**	87.29	**88.50**
Theis et al. (w/o attributes)	67.80	77.64	73.10	**85.77**	**76.97**	81.52	66.23	52.31	57.65	55.72	91.32	**89.60**
Theis et al. (w/ attributes)	66.25	64.23	65.21	76.16	72.52	73.56	65.12	47.69	63.51	56.47	82.35	87.24

Πίνακας 35 Μετρική ακρίβειας σε % για την πρόβλεψη επόμενης δραστηριότητας διαφορετικών δημοσιεύσεων. Το καλύτερο, το δεύτερο καλύτερο και το τρίτο καλύτερο σύστημα επισημαίνονται με κόκκινο, μπλε και πράσινο

Συγκρίνοντας λοιπόν τα δικά μου αποτελέσματα για τα σύνολα δεδομένων της αίτησης και της προσφοράς δανείου με τα αντίστοιχα αποτελέσματα των πιο πάνω δημοσιεύσεων για την αίτηση και προσφορά δανείου (BPI 2012 O , BPI 2012 A) παρατηρούμε ότι όχι μόνο είναι αρκετά ανταγωνιστικά αλλά πολλές φορές καλύτερα από τις άλλες προσεγγίσεις. Αρχίζοντας από το σύνολο δεδομένων offer, στην προσέγγιση της διπλωματικής μου κατάφερα να έχω μία ακρίβεια 81,6% η οποία είναι η τέταρτη καλύτερη. Αντίστοιχα στο σύνολο δεδομένων application η ακρίβεια που έχω 83.4% ξεπερνά τις υπόλοιπες προσεγγίσεις. Φυσικά τα δεδομένα δεν είναι τα ίδια και η σύγκριση αυτή αποσκοπεί στην ενστικτώδη αξιολόγηση των αποτελεσμάτων της διπλωματικής με τα αμέσως επόμενα δεδομένα τα οποία συνδέονται με τα δικά μου.

6

Επίλογος

6.1 Σύνοψη και συμπεράσματα

Σε αυτή τη διπλωματική εργασία προτείνουμε μία προσέγγιση βασισμένη στη βαθιά μάθηση για την πρόβλεψη της επόμενης δραστηριότητας σε μία επιχειρησιακή διεργασία. Τα δεδομένα που χρησιμοποιήθηκαν αφορούσαν αρχεία καταγραφής γεγονότων που λήφθηκαν από τον τραπεζικό τομέα και συγκεκριμένα από δεδομένα αιτήσεων και προσφορών δανείων ενός Ολλανδικού οικονομικού ινστιτούτου. Το μοντέλο πρόβλεψης βασίζεται σε ένα αναδρομικό νευρωνικό δίκτυο LSTM, επιτρέποντας την πρόβλεψη της επόμενης δραστηριότητας σε ένα ίχνος με είσοδο μία ή περισσότερες προηγούμενες δραστηριότητες.

Η προσέγγιση αποτελείται από τη φάση της προ-επεξεργασίας των δεδομένων, της υλοποίησης του νευρωνικού δικτύου LSTM και της εξαγωγής των αποτελεσμάτων. Κατά την προ-επεξεργασία έγινε η απομόνωση των ιχνών του αρχείου καταγραφής γεγονότων, η δημιουργία λεξικού ακεραίου μέσω της ordinal κωδικοποίησης, η κωδικοποίηση one-hot, η δημιουργία των διανυσμάτων εισόδου-εξόδου για την εκπαίδευση του νευρωνικού και τέλος ο διαχωρισμός των δεδομένων σε training/validation/test σύνολα. Για την υλοποίηση του νευρωνικού δικτύου έγινε αρχικά καθορισμός της αρχιτεκτονικής με επιλογή ορισμένων παραμέτρων όπως ο αλγόριθμος βελτιστοποίησης (optimizer), η συνάρτηση ενεργοποίησης (activation function) και η απώλεια (loss). Στην συνέχεια έγιναν αρκετά πειράματα για να επιλεχθούν οι βέλτιστες υπερ-παράμετροι για τον κατάλληλο αριθμό επιπέδων (layers), τον κατάλληλος αριθμός νευρώνων (neurons) και το ποσοστό του περιορισμού ενεργοποίησης (dropout). Αυτά τα πειράματα έγιναν για κάθε συνδυασμό στον αριθμό των δραστηριοτήτων εισόδου και εξόδου που επέλεξα να δοκιμάσω στην διπλωματική αυτή. Στη συνέχεια εκπαίδευσα το νευρωνικό δίκτυο και έγινε η εξαγωγή των αποτελεσμάτων και προβλέψεων για κάθε περίπτωση ξεχωριστά.

Τα αποτελέσματα, όπως παρουσιάστηκαν και στην προηγούμενη ενότητα υποδεικνύουν ότι η προτεινόμενη προσέγγιση είναι κατάλληλη για το πρόβλημα που καλείται να λύσει η διπλωματική εργασία. Η ακρίβεια του μοντέλου είναι συγκρίσιμη με αυτή άλλων προσεγγίσεων της βιβλιογραφίας ενώ σε ορισμένες περιπτώσεις ξεπερνά την ακρίβεια των δημοσιεύσεων που υπάρχουν για παρόμοια σύνολα δεδομένων.

6.2 Μελλοντικές επεκτάσεις

Το μοντέλο που αναπτύχθηκε στην παρούσα διπλωματική εργασία θα μπορούσε να βελτιωθεί και να επεκταθεί ως προς τις πιο κάτω κατευθύνσεις:

- Αξιολόγηση του μοντέλου σε άλλα σύνολα δεδομένων. Η αξιολόγηση μπορεί να γίνει τόσο σε παρόμοια σύνολα όπως το dataset από το Business Process Intelligence Challenge του 2012 (BPIC'12) για να γίνει και καλύτερη σύγκριση με τις προσεγγίσεις της βιβλιογραφίας, όσο και σε σύνολα δεδομένων από άλλους τομείς πχ. Νοσοκομεία, δημόσιους φορείς, βιομηχανίες κα.
- Ενσωμάτωση περαιτέρω χαρακτηριστικών του αρχείου καταγραφής γεγονότων στα δεδομένα. Για παράδειγμα κάποια από τα χαρακτηριστικά τα οποία είναι διαθέσιμα και μπορούν να δίνονται παράλληλα ως είσοδος στο LSTM είναι: ο σκοπός και ο τύπος του δανείου, η μηνιαία δόση και η διάρκεια αποπληρωμής, το credit score του πελάτη, το ποσό δανεισμού που ζήτησε ο πελάτης, το ποσό δανεισμού που προσφέρει η τράπεζα κτλπ. Ενσωματώνοντας αυτά τα δεδομένα στο νευρωνικό δίκτυο η ακρίβεια της πρόβλεψης τη διαδικασίας πιθανολογούμε ότι θα βελτιωθεί αρκετά.
- Πειραματισμός με άλλα μοντέλα μηχανικής μάθησης πέρα από το LSTM.
- Πρόβλεψη και άλλων χαρακτηριστικών του αρχείου καταγραφής γεγονότων όπως η χρονική διάρκεια ολοκλήρωσης των δραστηριοτήτων καθώς και ολόκληρου του trace, η κατανομή των πόρων, και η έκβαση του case.

7

Βιβλιογραφία

Van Der Aalst, W. M. P. (2013) 'Business Process Management: A Comprehensive Survey', *ISRN Software Engineering*. Hindawi Publishing Corporation, 2013, p. 37. doi: 10.1155/2013/507984.

van der Aalst, W. (2016) 'Data Science in Action', in *Process Mining*. Springer Berlin Heidelberg, pp. 3–23. doi: 10.1007/978-3-662-49851-4_1.

Van Der Aalst, W. (2012) 'Process mining: Overview and opportunities', *ACM Transactions on Management Information Systems*. doi: 10.1145/2229156.2229157.

Van Der Aalst, W. *et al.* (2012) 'Process mining manifesto', in *Lecture Notes in Business Information Processing*. Springer Verlag, pp. 169–194. doi: 10.1007/978-3-642-28108-2_19.

Baier, T., Mendling, J. and Weske, M. (2014) 'Bridging abstraction layers in process mining', *Information Systems*. Elsevier Ltd, 46, pp. 123–139. doi: 10.1016/j.is.2014.04.004.

Berti, A., van Zelst, S. J. and van der Aalst, W. (2019) 'Process Mining for Python (PM4Py): Bridging the Gap Between Process- and Data Science', *CEUR Workshop Proceedings*. CEUR-WS, 2374, pp. 13–16. Available at: http://arxiv.org/abs/1905.06169 (Accessed: 10 June 2021).

Ceci, M. *et al.* (2014) 'Completion Time and Next Activity Prediction of Processes Using Sequential Pattern Mining', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Verlag, pp. 49–61. doi: 10.1007/978-3-319-11812-3_5.

Chinces, D. and Salomie, I. (2015) 'Optimizing spaghetti process models', in *Proceedings - 2015 20th International Conference on Control Systems and Computer Science, CSCS 2015*. Institute of Electrical and Electronics Engineers Inc., pp. 506–511. doi: 10.1109/CSCS.2015.15.

Efrén *et al.* (2021) *Deep Learning for Predictive Business Process Monitoring: Review and Benchmark*. Available at: https://nextcloud.citius.usc.es/index.php/s/drMbTeGNKTE9axJ

(Accessed: 28 June 2021).

Evermann, J., Rehse, J.-R. and Fettke, P. (2016) 'Predicting Process Behaviour using Deep Learning', *Decision Support Systems*. Elsevier B.V., 100, pp. 129–140. doi: 10.1016/j.dss.2017.04.003.

Gupta, N. (2013) 'Network and Complex Systems Artificial Neural Network', 3(1). Available at: www.iiste.org (Accessed: 17 June 2021).

Hammer, M. *et al.* (1993) 'Reengineering the corporation: A manifesto for business revolution', *Business Horizons*. Elsevier, 36(5), pp. 90–91. Available at: https://econpapers.repec.org/RePEc:eee:bushor:v:36:y:1993:i:5:p:90-91 (Accessed: 9 May 2021).

Lakshmanan, G. T. *et al.* (2015) 'A markov prediction model for data-driven semi-structured business processes', *Knowledge and Information Systems*. Springer London, 42(1), pp. 97–126. doi: 10.1007/s10115-013-0697-8.

Manyika, J. *et al.* (2011) *Big data: The next frontier for innovation, competition, and productivity | Request PDF*. Available at: https://www.researchgate.net/publication/312596137_Big_data_The_next_frontier_for_innovation_competition_and_productivity (Accessed: 9 June 2021).

Navarin, N. *et al.* (2018) 'LSTM networks for data-aware remaining time prediction of business process instances', in *2017 IEEE Symposium Series on Computational Intelligence, SSCI 2017 - Proceedings*. Institute of Electrical and Electronics Engineers Inc., pp. 1–7. doi: 10.1109/SSCI.2017.8285184.

Tax, N. *et al.* (2017) *Predictive Business Process Monitoring with LSTM Neural Networks*.

Tello-Leal, E. *et al.* (2018) 'Predicting activities in business processes with LSTM recurrent neural networks', in *10th ITU Academic Conference Kaleidoscope: Machine Learning for a 5G Future, ITU K 2018*. Institute of Electrical and Electronics Engineers Inc. doi: 10.23919/ITU-WT.2018.8598069.

Thomas H. Davenport (1993) 'Process Innovation -- Reengineering Work Through Information Technology, By Thomas H. Davenport, Harvard Business School Press, 1993, p. 326, Price £29.95 ISBN 0 87584 366 2', *R&D Management*. Wiley, 25(4), pp. 421–421. doi: 10.1111/j.1467-9310.1995.tb01348.x.

Weske, M. (2012) *Business process management: Concepts, languages, architectures*, *Business Process Management: Concepts, Languages, Architectures, Second Edition*. Springer Berlin Heidelberg. doi: 10.1007/978-3-642-28616-2.