Ένα από τα κύρια χαρακτηριστικά της εποχής μας είναι τα δεδομένα μεγάλου όγκου. Πλέον είναι εφικτό να συλλέγουμε πληροφορίες από πληθώρα πηγών: τα έξυπνα κινητά, και γενικότερα οι έξυπνες συσκευές, τα κοινωνικά δίκτυα και το σύστημα υγείας αποτελούν ένα μικρό τμήμα των δυνητικών πηγών δεδομένων. Με την αύξηση των δεδομένων προέκυψαν δύο σημαντικά προβλήματα: η αποθήκευση και η επεξεργασία τους. Προκειμένου να ξεπεραστεί το πρώτο πρόβλημα αναπτύχθηκαν τεχνικές οι οποίες επιτρέπουν τόσο την γρήγορη όσο και την αξιόπιστη αποθήκευση και αναζήτηση της πληροφορίας. Για το δεύτερο πρόβλημα αναπτύχθηκαν καινούρια προγραμματιστικά πλαίσια *(frameworks)* τα οποία επιτρέπουν την επεξεργασία των δεδομένων χρησιμοποιώντας συστάδες *(clusters)* υπολογιστών.

Στην παρούσα διπλωματική εργασία, χρησιμοποιείται το προγραμματιστικό πλαίσιο *Apache Spark* το οποίο επιτρέπει την παράλληλη επεξεργασία δεδομένων. Για την υλοποίηση επιλέχθηκε η δωρεάν έκδοση του *Databricks (Databricks community edition)* η οποία παρέχει χώρο αποθήκευσης δεδομένων και διαθέσιμους πόρους για την επεξεργασία τους. Πραγματοποιήθηκαν δύο τύποι αναλύσεων: ανάλυση κατηγοριοποίησης *(classification)* και συνεργατικού φιλτραρίσματος *(collaborative filtering)*.

Στην ανάλυση κατηγοριοποίησης χρησιμοποιήθηκαν δύο σύνολα δεδομένων, ένα δυαδικό και ένα πολλαπλών κλάσεων εξόδου, στα οποία εφαρμόστηκε μία σειρά τεχνικών κατηγοριοποίησης, με σκοπό να συγκρίνουμε τις διάφορες τεχνικές κατηγοριοποίησης, να εξεταστεί η επεκτασιμότητα κάθε αλγορίθμου αλλά και να εξαχθούν συμπεράσματα ως προς την επίδραση των παραμέτρων της εκάστοτε τεχνικής. Για τις αναλύσεις χρησιμοποιήθηκε η βιβλιοθήκη *Spark MlLib,* η οποία παρέχεται από το *framework Apache Spark* και περιλαμβάνει υλοποιήσεις τεχνικών μηχανικής μάθησης βελτιστοποιημένες για κατανεμημένο περιβάλλον. Εκτός της τυπικής μεθοδολογίας κατηγοριοποίησης, πραγματοποιήθηκε μία σειρά αναλύσεων κατηγοριοποίησης δύο βημάτων, όπου στο πρώτο βήμα χρησιμοποιήθηκε μία αυτόματη μέθοδος για την εύρεση ενός υποσυνόλου των χαρακτηριστικών εισόδου και στη συνέχεια, χρησιμοποιώντας αυτό το υποσύνολο, επαναλήφθηκε η διαδικασία κατηγοριοποίησης. Σκοπός αυτής της διαδικασίας ήταν να μελετηθεί η επίδραση του αριθμού των χαρακτηριστικών τόσο στην ποιότητα των αποτελεσμάτων όσο και στον χρόνο εκτέλεσης.

Στην ανάλυση *collaborative filtering* χρησιμοποιήθηκε ένα πραγματικό σύνολο δεδομένων που περιλαμβάνει βαθμολογήσεις ταινιών από χρήστες. Με βάση αυτό εξετάστηκε η επίδραση διαφόρων παραμέτρων του αλγορίθμου στην ποιότητα των τελικών αποτελεσμάτων αλλά και του χρόνου εκτέλεσής του. Για την υλοποίηση της διαδικασίας χρησιμοποιήθηκε και πάλι η βιβλιοθήκη *Spark MlLib*, ενώ επιπλέον εξήχθησαν προβλέψεις και για ένα νέο χρήστη.