ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Χανιωτάκης Ευάγγελος
Μεταπτυχιακός Φοιτητής

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτυχιακής Εργασίας: Καθηγητής, Δ. Πλεξουσάκης

Μ. Τσικνάκης (επιβλέπων)

Τετάρτη, 3 Μαρτίου 2021 , ώρα 10:00 π.μ.

Join Zoom Meeting

https://zoom.us/j/96552534995

"Μια Επεκτάσιμη Πλατφόρμα Επιστήμης Δεδομένων, βασισμένη σε Τεχνολογίες Ανοιχτού Κώδικα με Εφαρμογή Προγνωστικής Ανάλυσης για τη νόσο του Συνδρόμου Οξείας Αναπνευστικής Δυσχέρειας"

Περίληψη

Η συνεχής ανάπτυξη μεγάλου όγκου βιοϊατρικών δεδομένων στην υγειονομική περίθαλψη δημιουργεί σημαντικές προκλήσεις για την αποτελεσματική τους διαχείριση. Αυτή η ανάγκη έκανε αναπόφευκτη την υιοθέτηση μεγάλων υποδομών δεδομένων και σχετικών τεχνικών από οργανισμούς υγειονομικής περίθαλψης, προκειμένου να εξερευνήσουν αποτελεσματικά τον πλούτο των δεδομένων του πραγματικού κόσμου που δημιουργούνται με στόχο τη βελτίωση της ποιότητας των υπηρεσιών υγείας. Στη βιομηχανία υγειονομικής περίθαλψης, υπάρχουν διάφορες μεγάλες πηγές δεδομένων, που χαρακτηρίζονται από ετερογένεια. Αυτές περιλαμβάνουν νοσοκομειακά συστήματα πληροφοριών (HIS) και ιατρικά αρχεία ασθενών (EHRs), αποτελέσματα εργαστηριακών διαδικασιών και εξετάσεων που βρίσκονται σε σχετικά συστήματα πληροφοριών (Laboratory Information Systems - LIS), δεδομένα από συνεχή παρακολούθηση ασθενών (π.χ. σε μία μονάδα εντατικής θεραπείας - ΜΕΘ) και δεδομένα από έξυπνες συσκευές,

όπως φορητά. Επίσης, πολύ μεγάλα σύνολα δεδομένων δημιουργούνται από κλινικές και ερευνητικές εργασίες που σχετίζονται με τη γονιδιωματική. Όσον αφορά τη γονιδιωματική, ο ρυθμός ανάπτυξης κατά την τελευταία δεκαετία ήταν επίσης πραγματικά εκπληκτικός, με τον συνολικό αριθμό δεδομένων αλληλούχισης που παράγονται να διπλασιάζεται περίπου κάθε επτά μήνες. Αυτά τα δεδομένα απαιτούν αποτελεσματική διαχείριση και ανάλυση προκειμένου να εξάγουν ουσιαστικές και εφαρμόσιμες πληροφορίες.

Κατά την ανάπτυξη τέτοιων λύσεων πρέπει να αντιμετωπιστεί μια σειρά από προκλήσεις και επιπλοκές που συνδέονται με κάθε βήμα του σχεδιασμού συστημάτων για την διαχείριση τέτοιων μεγάλων συνόλων δεδομένων υγειονομικής περίθαλψης. Αυτές μπορούν να επιλυθούν μόνο χρησιμοποιώντας υψηλής ποιότητας υπολογιστικές λύσεις για ανάλυση μεγάλων δεδομένων. Ειδικά στην τρέχουσα κατάσταση της πανδημίας COVID-19, οι επιπλοκές που μπορεί να εμφανιστούν μετά την έναρξη αυτής της ασθένειας στη ζωή του ανθρώπου είναι πραγματικά σημαντικές. Μια σημαντική τέτοια επιπλοκή είναι το σύνδρομο οξείας αναπνευστικής δυσχέρειας (ARDS), το οποίο είναι μια σοβαρή αναπνευστική κατάσταση με υψηλή θνησιμότητα και σχετική νοσηρότητα. Ένας μεγάλος αριθμός βασικών και κλινικών μελετών έχουν δείξει ότι η έγκαιρη διάγνωση και παρέμβαση είναι καθοριστικής σημασίας για τη βελτίωση του ποσοστού επιβίωσης των ασθενών με ARDS. Επομένως, υπάρχει επιτακτική ανάγκη για την ανάπτυξη και κλινική δοκιμή προγνωστικών μοντέλων για συμβάντα ARDS, τα οποία θα μπορούσαν να βελτιώσουν την κλινική διάγνωση ή τη διαχείριση του ARDS.

Στην παρούσα διατριβή, εστιάσαμε σε δύο διαφορετικούς στόχους: συγκεκριμένα α) να σχεδιάσουμε μια επεκτάσιμη πλατφόρμα διαχείρισης μεγάλου όγκου δεδομένων, βασισμένοι σε τεχνολογίες ανοιχτού κώδικα, και β) να εκμεταλλευτούμε την πλατφόρμα και δημόσια διαθέσιμα μεγάλα σύνολα κλινικών δεδομένων προκειμένου να αναπτύξουμε μοντέλα μηχανικής μάθησης για την πρόβλεψη συμβάντων οξείας αναπνευστικής δυσχέρειας (ARDS) μέσω κοινώς διαθέσιμων παραμέτρων, συμπεριλαμβανομένων των βασικών χαρακτηριστικών και των κλινικών και εργαστηριακών παραμέτρων.

Η διατριβή χωρίζεται σε δύο κύρια μέρη. Το πρώτο μέρος παρουσιάζει και αναλύει λεπτομερώς όλες τις διαδικασίες, τα υλικά και τις μεθόδους που υιοθετήθηκαν για την ανάπτυξη αυτής της πλατφόρμας διαχείρισης μεγάλων δεδομένων. Εστιάσαμε στις επιπλοκές και τις δυσκολίες που προκύπτουν κατά τη δημιουργία και τη χρήση τέτοιων συστημάτων σε μεγάλα βιοϊατρικά δεδομένα, όπως το σύνολο δεδομένων MIMIC-III. Το δεύτερο μέρος αυτής της διατριβής, περιγράφει τον τρόπο με τον οποίο χειριστήκαμε αυτήν την κλινική βάση δεδομένων, για να πραγματοποιήσουμε μια μελέτη αξιολόγησης

της πλατφόρμας μας, σε ένα πραγματικό κλινικό σενάριο για το ARDS. Ο στόχος της μελέτης μας ήταν να αναπτύξουμε και να αξιολογήσουμε μια νέα εφαρμογή αλγοριθμικών μοντέλων, Random Forest και Logistic Regression, που εκπαιδεύτηκαν σε δεδομένα σχετικά με την υγεία των ασθενών, για την πρώιμη διάγνωση και πρόβλεψη του ARDS. Η προσέγγιση μας επιτυγχάνει καλύτερα αποτελέσματα σε όλες τις μετρήσεις, σε σύγκριση με σχετικές δημοσιευμένες προσπάθειες που επίσης χρησιμοποιούν τη βάση δεδομένων MIMIC III για την ανάπτυξη προγνωστικών μοντέλων για ARDS. Συγκεκριμένα, και τα δύο αλγοριθμικά μοντέλα μας έχουν καλύτερη απόδοση στην πρόβλεψη ARDS, με κυρίαρχο το Random Forest με 10-fold cross validation, σύμφωνα με την περιοχή κάτω από την καμπύλη AUC (95,1%), την ακρίβεια (98,0%), την ειδικότητα (98,62%) και την ευαισθησία (96,25%).

University of Crete

Computer Science Department

M.Sc. Thesis presentation / examination

Chaniotakis Evangelos

Master's Thesis Supervisor: Professor, D. Pleksousakis

M. Tsiknakis (Thesis CO-Advisor)

Wednesday, 3 March 2021, 10:00 a.m

Join Zoom Meeting

https://zoom.us/j/96552534995

"A Scalable Data Science Platform built on Open Source Technologies with Application of Predictive Analytics on Acute Respiratory Distress Syndrome disease"

Abstract

The continuous growth of high volumes of biomedical data in healthcare generates significant challenges for their efficient management. This need has made inevitable the

adoption of big data infrastructures and relevant techniques from healthcare organizations, in order for them to efficiently explore the wealth of real-world data generated with the objective to improve the quality of healthcare services. In the healthcare industry, various big data sources, that are characterized by heterogeneity, exist. These include hospital information systems (HIS) and medical records of patients (EHRs), results of laboratory procedures and examinations residing in relevant information systems (Laboratory Information Systems - LIS), data from continuous patient monitoring (e.g. in an Intensive Care Unit - ICU) and data from smart devices, such as wearables. Also, very big data sets are generated from genomics-related clinical and research work. Regarding genomics, the rate of growth over the last decade has also been truly astonishing, with the total amount of sequence data produced doubling approximately every seven months. This data requires efficient management and analysis in order to derive meaningful and actionable information.

In developing such solutions, a range of challenges and complications associated with each step of the pipeline for handling such healthcare big data sets need to be addressed. These can only be resolved by using high-quality computing solutions for big data analysis. Especially in the current situation of the COVID-19 pandemic, complications that might occur after the onset of this disease are really important. An important such complication is Acute Respiratory Distress Syndrome (ARDS), which is a serious respiratory condition with high mortality and associated morbidity. A large number of basic and clinical studies have demonstrated that early diagnosis and intervention are key to improving the survival rate of patients with ARDS. Therefore, there is a pressing need for the development and clinical testing of predictive models for ARDS events, which might improve the clinical diagnosis or the management of ARDS.

In the present thesis, we focused on two distinct objectives; namely a) to design a scalable data science platform, built on open source technologies, and b) to exploit the platform and publically available big healthcare datasets to develop machine learning models for predicting acute respiratory distress syndrome (ARDS) events through commonly available parameters, including baseline characteristics and clinical and laboratory parameters.

This thesis is divided into two main parts. The first part presents and analyzes in detail all the procedures, materials, and methods adopted to develop this big data management platform. We report on the complications and difficulties that arise in creating and using such systems with large biomedical datasets, such as the MIMIC-III dataset. The second part of the thesis describes how we exploit this clinical database, to perform an evaluation study of our platform on a real world clinical scenario for ARDS. The objective of the study

was to develop and evaluate a novel application of machine learning models for predicting acute respiratory distress syndrome (ARDS. We employ random forests and logistic regression algorithmic models, trained on patient health record data for the early prediction and diagnosis of ARDS. Our approach achieves better results in all metrics that are based on AUC, when compared to relevant published efforts using the MIMIC III dataset to develop predictive models of ARDS. Specifically, both of our algorithmic models outperform in ARDS prediction, with 10-fold cross validated Random Forest being dominant, according to AUC (95.1%), Accuracy (98.0%), Specificity (98.62%) and Sensitivity (96.25%).