INNODOCT 2020
Valencia, 11th-13th November 2020
DOI: http://dx.doi.org/10.4995/INN2020.2020.11890

Project based learning in Biomedical Data Science using the MIMIC III open dataset

Luis Alcalá[a], Juan M García-Gómez[b,c], Carlos Sáez[b,d]

[a]Grado en Ingeniería Biomédica, Escuela Técnica Superior de Ingeniería Industrial, Universitat Politècnica de Valencia, España, luialpe@etsii.upv.es, [b]Intituto Universitario de Tecnologías de la Información y Comunicaciones, Universitat Politècnica de Valencia, España, [c]juanmig@upv.es, [d]carsaesi@upv.es

Abstract

The subjects Health Information Systems and Telemedicine and Data Quality and Interoperability of the Degree and Master in Biomedical Engineering of the Universitat Politècnica de València, Spain, address learning outcomes related to managing and processing biomedical databases, using health information standards for data capture and exchange, data quality assessment, and developing machine-learning models from these data. These learning outcomes cover a large range of distinct activities in the biomedical data life-cycle, what may hinder the learning process in the limited time assigned for the subject. We propose a project based learning approach addressing the full life-cycle of biomedical data on the MIMIC-III (Medical Information Mart for Intensive Care III) Open Dataset, a freely accessible database comprising information relating to patients admitted to critical care units. By means of this active learning approach, students can achieve all the learning outcomes of the subject in an integrated manner: understanding the MIMIC-III data model, using health information standards such as International Classification of Diseases 9th Edition (ICD-9), mapping to interoperability standards, querying data, creating data tables and addressing data quality towards applying reliable statistical and machine learning analysis and, developing predictive models for several tasks such as predicting in-hospital mortality. MIMIC-III is widely used in the academia and science, with a large amount of publicly available resources and scientific articles to support the students learning. Additionally, the students will gain new competences in the use of Open Data and Research Ethics and Compliance Training.

Keywords: *Project based learning, biomedical engineering, MIMIC, data model, mortality prediction*

1. Introduction

Information Systems and Telemedicine (SIT) and Data Quality and Interoperability (DQI) are two subjects of the Degree in Biomedical Engineering and Master in Biomedical Engineering, respectively, of the Universitat Politècnica de València, Spain. The main learning goals of SIT include the processing of electronic health records and the development of machine learning and decision support systems using biomedical data. On the other hand, the main learning goals of DQI include describing the quality of biomedical data and assessing their consequent problems on data analysis, such as those learnt in SIT. The correct use of health data in data science, from statistical analysis to machine learning development is essential for an efficient healthcare system and, especially, for improving the patients wellbeing.

Part of the training in these subjects is related to understanding the problems of data quality and data pre-processing, as well as being trained in specific prediction and machine-learning models. Learning these methods needs a theoretical background which may result too abstract for the student without any practical open examples. We consider therefore that having an approach of a project based learning (Blumenfeld et al., 1991) which runs through all the data science life-cycle stages, from pre-processing, data quality analyses, to the development of models may result the on the most useful approach for the student to acquire the needed competencies. Furthermore, the use of real and tangible data and the methods to process it is a novel experience for the students, which can bring them a wider, more real perspective about biomedical data science. Currently, the DQI subject is supported by its own project based learning approach (Sáez, Mañas, Muñoz-soler, & García-, 2017), however, the complementariness of the two subjects can allow a within-subject approach which can boost the learning process using a real world casuistic as provided by the MIMIMC III data .In this work, we propose a new project based learning approach for the SIT and DQI subjects, which addresses the full life-cycle of biomedical data on the MIMIC-III (Medical Information Mart for Intensive Care III) Open Dataset, a freely accessible database (Johnson et al., 2016) comprising information relating to patients admitted to critical care units.

2. The MIMIC Project

The MIMIC project (Moody et al., 1996) was published in 1999 by the Computational Physiology Laboratory of the Massachusetts Institute of Technology (MIT), staring data from 90 patients from the Intensive Care Unit (ICU) of the Beth Israel Deaconess Medical Center, Boston. Since then, several versions have been developed and published, staring more and more data from an increasing number of patients.

In this project, we use the latest available version (Johnson et al., 2016), the MIMIC III v1.4, published in 2016. It is composed of two different databases, a waveform database, and a clinical database (CDB). For the purpose of this project, only the clinical database is needed. The CDB is formed with 26 related tables that collect all the information needed, making a total of 43.3 GB. The data comes from 53 423 distinct hospital admissions related to ICU, covering 38 597 distinct adult patients and 7870 neonates, from 2001 to 2012. The CDB features clinical deidentified data coming from a wide variety of areas, from laboratory tests to International Classification of Diseases, Ninth Revision (ICD-9) codes ("ICD - ICD-9 - International Classification of Diseases, Ninth Revision," n.d.), as seen in Figure 1. Three principal sources were considered to include the data:

- Critical care information systems (Hospital-ICU)
- Hospital electronic health record databases (Hospital)
- Social Security Administration Death Master File (External)

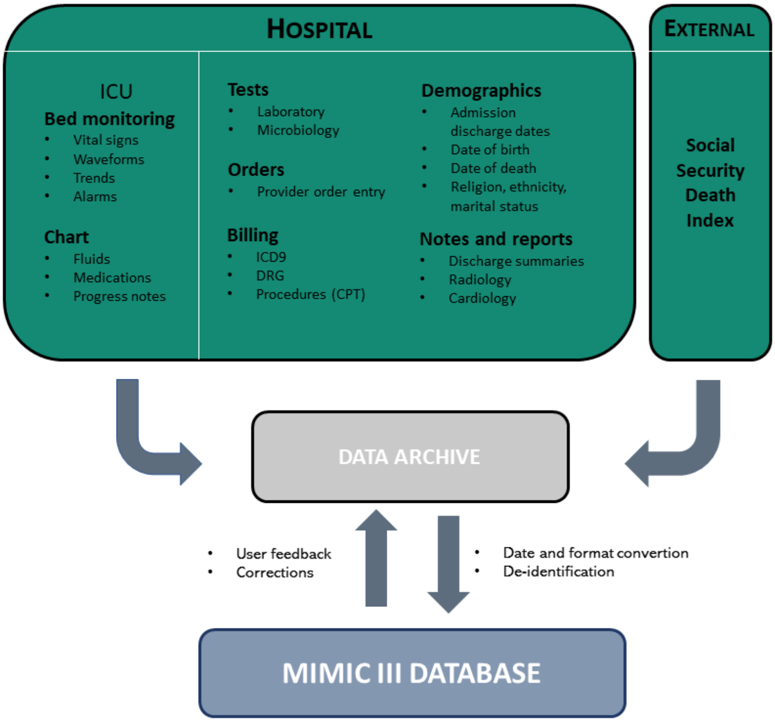

Fig. 1 Overview of the MIMIC III database

The data present on the 26 tables has some specific characteristics given the deidentifying process applied by the dataset creators. For instance, every date of each of the admissions into the hospital has being randomized into the future, in a consisting manner for each patient, resulting in dates that take place between the years 2100 and 2200. Nevertheless, an approximate time of the day, day of the week and seasonality have been conserved. On the other hand, patients aged over 89 had their dates of birth shifted. This results in patients with ages of over 300 years.

MIMIC III v1.4 is a freely accessible database, however, the completion of Human Research Data or Specimens Only Research Basic Course (CITI Training) under the requirements of the MIT is needed to be granted the access. This course has a length of approximately 6 hours.

Once the access to the database is granted, and due to the large amount of data present on the tables, the need of database management software is required, such as PostgreSQL and Google BigQuery. Data treatment and analysis will be done both in SQL querys and in R language, prediction models will be developed in R language.

3. Methods

The project based learning approach covers the different stages of data analysis, with the ultimate goal of developing predictive models. In our case, the target tasks included the prediction of in hospital mortality, equivalently to tasks that have already addressed in the scientific literature using the MIMIC III data (Johnson, Pollard, & Mark, 2017). Our proposal is that students replicate the work we describe next.

Table 1. Mains tables used to extract data for model generation

Table Name	Size (GB)	Number of variables	Description
ADMISSIONS	0.012	19	Hospital admissions associated with an IC stay
CHARTEVENTS	34.5	15	Events occurring on a patient chart
DIAGNOSES_ICD	0.019	5	Diagnoses relating to a hospital admissior coded using the ICD9 system
ICUSTAYS	0.006	12	List of ICU admissions
PATIENTS	0.003	8	Patients associated with an admission to ICU

With the amount of information available in the database, the first step was to clarify which data was valid and useful for these prediction goals. The full CDB was a non – viable approach, due to its magnitude and the hardware necessary to process it, which not every student has access to. We decided then to extract only the necssary data from specific tables (Table 1) creating two compacted datasets that only integrate the useful data for the project. The first dataset (Dataset A) included physiological and administrative data used to train models such as K_Nearest_Neighbours (KNN), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA). The second one (Dataset B) included processed data from ICD-9 codes, available in the DIAGNOSES_ICD table, training models such as Random Forest (RF) or Gradient Boosting (GB).

For the purpose of this study, each admission, even readmissions of the same patient, was treated as an individual measure. Nevertheless, we made sure that, in training and validation sets, the same patients were in the same group, to avoid misleading results. Each patient has an unique identifier (SUBJECT_ID), an unique admission identifier (HADM_ID) and an unique ICU stay identifier (ICUSTAY_ID).

Once the CDB was downloaded and installed in a local PostgreSQL server, several querys were performed on the CHARTEVENTS table. As seen in Table 1, this dataset is the main problem size ways, which forced us to use SQL language to extract the important data.

A cohort for our two datasets was defined based on our goal's interests. From over 50000 admissions available in the dataset, several inclusion criteria were adopted to properly reduce the size of valid admissions. Firstly, we removed non-adult related admissions, more specifically ICU admissions of patients aged under 16 years old. Secondly, stays under four hours were as well removed due to the non-interest of using these data in mortality prediction models. Lastly, we removed admissions related to organ donor accounts.

Regarding Dataset A, we defined the window of time to extract the time series of vital signs at the CHARTEVENTS table. This window began at ICU admission and ended up to around 24 hours after ICU admission. The vital signs measurements selected for the study, based on what was previously studied in the literature (Johnson et al., 2017), are heart rate, systolic, diastolic and mean blood pressure, respiration rate, temperature (ºC), oxygen saturation and glucose. Measurements were made to select key values of the time series such as minimum, maximum, mean and standard deviation. At the same time, non-vital signs information such as age, weight, gender, the admissions identifier and the patient identifier, was included in the dataset.

For Dataset B, cohort criteria remained the same, but, due to the size of the DIAGNOSES-ICD table, data processing was made on R Studio. Several ICD-9 codes were related with each patient admissions. Firstly, we associated each of those codes to the corresponding ICD-9 chapter and the we transformed that data into dummy value, creating a new dataset featuring the admission identifier for the patient and the dummy values corresponding to the 19 possible ICD-9 chapters.

Once we had both datasets ready, and for the both of them a training set with 70% of the data and a validation, with the remaining 30%, were created.

A Principal Component Analysis (PCA) was made for dataset A, deriving from it a new dataset with the 8 first components. This dataset was then used to train three different models: Lineal Discriminative Analysis, Quadratic Discriminative Analysis and K_Neighbours.

As said before, dataset B was used to train two different and more advanced models, Random Forest and Gradient Boosting.

For both datasets, we calculated data quality dimensions (Sáez, Martínez-Miranda, Robles, & García-Gómez, 2012) including completeness, consistency, and correctness. Other dimensions including temporal and multi-source stability, and contextualization, were left for further work.

4. Results

4.1 Models

Results for the mortality prediction models trained for both datasets are shown in Table 2.

Table 2. Model results for datasets A and B

	Model	Accuraccy
Dataset A	Lineal Discriminative Analysis	0.68
	Quadratic Discriminative Analysis	0.63
	K_Nearest_Neighbours	0.76
Dataset B	Random Forest	0.89
	Gradient Boosting	0.91

For dataset A, KNN showed the best results, having the best accuracy and the best ratio of patients predicted deceased vs patients deceased. Nevertheless, results are not as expected and further study of the CHARTEVENTS table is required to extract more reliable data. Regarding dataset B, results improve substantially for both models.

4.2 Project based learning

The third and most extense didactic unit (DU) of the SIT subject, addresses the topic of Clinical Decision Support Systems (CDDS), which encompasses the whole process of data processing and usage to train and develop the predictive models focused in CDDS. Previously, DU 1 relates to handling databases and DU 2 relates with working and understanding specific biomedical data terminologies, such and ICD-9 or DRG codes, as available in the MIMIC-III data.

Furthermore, DUs 1 and 2 of DQI cover the description and understanding of data quality dimensions such as completeness, consistency, correctness, contextualization, temporal and multisource stability.

Table 3. SIT and DQI didactic units covered by the project

Didactic Units	Task developed by the MIMIC III project	Specific learning goal covered
SIT - DU 1: Organization of Health Information Systems	Understanding and querying MIMIC III data, creating databases for prefictive modelling	Manging and using hospital databases
SIT - DU 2: Electronic Health Record Standards	Using the ICD-9 standard	Describing and using HER standards
SIT - DU 3: Clinical Decision Support Systems	Developing in-hospital mortality predictive models using distict mahine learning algorithms	Develop predictive models and CDSS using machine learning algorithms
DQI – DU 1: Introduction to Data Quality and Data Quality Dimensions	Describing de DQ dimensions that can be addressed in the MIMIC III dataset	Describing the different DQQ aspects of data and classifying them on dimensions
DQI – DU 2: Data quality Dimensions	Measuring and fixing completeness, consistency, correctness, contextualization, and temporal and multi-source stability dimensions, towards reliable predictive modelling	Measuring DQ dimensions and curating data

As seen in Table 3, the usage of the data given by the MIMIC project captures the essence of the most important SIT and DQI didactic units. Not only we could be able to show theorical examples based on this data, but to create a semi-guided final subject task that follows the path of the project described in this work and is able to show the students the real difficulties and the right methods to treat real tangible data.

Regarding the UPV competencies in Biomedical engineering, this project cover principally: 40(ES) Capacity for self-learning, consolidation and updating of new knowledge in the area of biomedical engineering, and for undertaking subsequent studies with a high degree of autonomy. 43(GE) Ability to learn new techniques and tools for analysis, modelling, design and optimisation. 5(ES) Possess knowledge of computer tools for analysing, calculating, visualising, representing and obtaining the necessary information to support analysis, calculation, design, development and management tasks related to biomedical engineering. 8(ES) Ability to integrate multidisciplinary knowledge associated with engineering, biology and medicine. 11(ES) Be able to understand the technical and functional characteristics of systems, methods and procedures used in prevention, diagnosis, therapy and rehabilitation.

5. Discussion

Our approach for a project based learning using the MIMIC III open dataset has been successful in this proposal stage, having demonstrated the MIMIC III data allows addressing the expected learning outcomes. The relathionship of part of SIT and DQI subjects results highly complementary and may allow for the students to boost the adquisition of the needed competences when working with the proposed project based learning approach through the Degree to the Master courses. The next step would be to test this project based learning with the students, study the effect it has on their skills, awareness and interest for the topic, and get their feedback to improve our methods.

One of the limitations we have found is that the anonymization of data regarding admission and birth dates can limit the applicability of longitudinal temporal variability analyses, as part of a data quality analysis (Sáez, C., Zurriaga, O., Pérez-Panadés, J., Melchor, I., Robles, M., & Garcia-Gomez, J. M., 2016). However, the seasonality effects can still be explored.In further work we will also study the multi-source variability analyses, using the included sources of data. Another limitation, is that the required CITI training course is a must do for the students of the SIT and DQI subject to be able to use MIMIC III data. However, this can be also considered an profitable situation, given that the learning goals of that additional course are extremely important for the future student’s competence in processing and analysing biomedical data.

To our knowledge, this is the first university cycle project based learning approach using the MIMIC III database covering the full biomedical data life-cycle towards data quality and predictive modelling. We only found the Coursera course "Clinical Data Models and Data Quality Assessments" ("Clinical Data Models and Data Quality Assessments | Coursera," n.d.) using the MIMIC III dataset to teach about data models and specific aspects of data quality. A course which our students might complementaryly take to improve their understanding on the dataset and data quality topic.

6. Conclusion

The primary objective of this work was to define a global project based learning for both SIT and DQI subjects. The proposed approach was defined, showing coverage of the subject learning goals, and ready to be further evaluated with students This project has addiotionally lead us to the creation of a new array of data based on the MIMIC III database for the scholarly use supporting this project based learning approach. The MIMIC III v1.4 dataset not only provides with an immense amount of clinical data, but is well prepare for the extraction and analysis of it. Processing the data present on the dataset A and B that we have developed, as well as creating multiple prediction mortality models and analysing the quality of it may be an enormous improvement of the learning curve of the students in this matters.

References

Blumenfeld, P. C., Soloway, E., Marx, R. W., Krajcik, J. S., Guzdial, M., & Palincsar, A. (1991). Motivating Project-Based Learning: Sustaining the Doing, Supporting the Learning. Educational Psychologist, 26(3–4), 369–398. https://doi.org/10.1080/00461520.1991.9653139

Clinical Data Models and Data Quality Assessments | Coursera. (n.d.). Retrieved June 15, 2020, from https://www.coursera.org/learn/clinical-data-models-and-data-quality-assessments

ICD - ICD-9 - International Classification of Diseases, Ninth Revision. (n.d.). Retrieved June 15, 2020, from https://www.cdc.gov/nchs/icd/icd9.htm

Johnson, A. E. W., Pollard, T. J., & Mark, R. G. (2017). Reproducibility in critical care: a mortality prediction case study. Proc. Mach. Learn. Res., 68, 361–376.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., … Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 1–9. https://doi.org/10.1038/sdata.2016.35

Moody GB, Mark RG. A Database to Support Development and Evaluation of Intelligent Intensive Care Monitoring. Computers in Cardiology 23:657–660 (1996)

Sáez, C., Mañas, A., Muñoz-soler, V., & García-, J. M. (2017). Project-based learning based on a national pilot project for the data quality control and standardization of maternal and child information applied to Biomedical Engineering University teaching, (October), 75–85.

Sáez, C., Martínez-Miranda, J., Robles, M., & García-GóMez, J. M. (2012). Organizing data quality assessment of shifting biomedical data. Studies in Health Technology and Informatics, 180, 721–725. https://doi.org/10.3233/978-1-61499-101-4-721

Sáez, C., Zurriaga, O., Pérez-Panadés, J., Melchor, I., Robles, M., & Garcia-Gomez, J. M. (2016). Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in Spain: a systematic approach to quality control of repositories. Journal of the American Medical Informatics Association, 23(6), 1085-1095.