DATA MINING CA1 (B9DA103)

Submitted To: Terri Hoare

Submitted By: Ashwin Ramdas
10534162

TABLE OF CONTENT

Chapter 1- CRISP-DM Critique

Chapter 2 - Critical Analysis of Big Data in Smart Farming

1. INTRODUCTION

Here, We critique CRISP DM and its techniques in terms of its application in the journals "Scalable Predictive Analysis in Critically Ill Patients Using a Visual Open Data Analysis Platform" by Van Poucke S, Zhang Z, Schmitz M and "Progression analysis of signals: Extending CRISP-DM to stream analytics" by P. Kalgotra and R. Sharda, on the basis of "The CRISP-DM Model: The New Blueprint for Data Mining" Shearer 2000.

2. SUMMARY

The *Cross-industry standard process for data mining (CRISP-DM)* was put forward by the people from Daimler, Teradata, SPSS, OHRA, and NCR in 1996 to create a standard data mining process across industries. Six major iterative phases are described by CRISP-DM, where every one of it had its own said tasks and deliverables such as documents and report. (CRISP-DM Guide.)

1. **Business Understanding:**

   The project goals and requirements is understood in the initial first phase, and a data mining problem definition is found from the knowledge gotten before, and designs a preliminary project plan .

   Different steps are-

   1. determining goals of the business- In this phase, we choose what the business is and what goals does it wish to achieve in a said timeframe;
   2. situation assessment- This step involves assessing the situation, the business is in right now and how does it work towards attaining its goals;
   3. data mining needs are determined- Here, we assess what data mining is used business for and data mining features are ascertained;
   4. project plan is produced-In this very important step, the final project plan is completed to meet the business needs
2. **Data Understanding:**

   Firstly, data collection is done in the data understanding phase after which familiarity of data is increased by per forming such tasks as, data quality troubles are identified, insights into the information collected is discovered, or interesting subsets to parse for hidden information are discovered. The link between the first two steps of the process are very close. The data mining challenges are formulated, and some understanding of the available information is required.

   Each different step for this stage is-

   1. Data collection- We collect data in this step, which is then stored in the database;
   2. Data description- Here, we correctly describe the data, which include checking if all the labels are correct, so that the data science algorithms can run smoothly;
   3. Data exploration- In this step, the data is explored to find what sort of data, we're dealing with;
   4. data quality verification- in here, we verify the quality of the information obtained
3. **Data Preparation** :

The final dataset is constructed after the data preparation phase, because it covers all the activities (the modeling tool(s)) is fed initial raw data. Tabling, recording, and
attribute selecting, data cleansing, new attributes development, and data transformation with modeling tools.

Different steps for this stage's implementation are-

1. Data selection- In this step we select the data we need to work on, for this we check the values which will have the most impact on the target variables and ignore the ones with the least impact;
2. Data cleaning- In this step, the data is cleaned to remove any null values or replace them, to clear out the unimportant columns of the dataset;
3. Data construction- Here, the data is constructed in such a way that it is optimal for the algorithms to work on;
4. data integration- the data is integrated to be fed into the model;
5. data formatting- the information is formatted

4. **Modelling:**

In this phase, the selection of the one among the various modeling techniques is performed and applied, and optimal values of their parameters are found. Usually, techniques for the similar data mining problem type are numerous. Specific data formats are required by some specific questions. There is a close link between Data Preparation and Modeling.

Different steps are-

1. modelling technique selection- Here, by looking at the data, we ascertain which of the modelling technique will we be using in our project;
2. test design generation- here, a test design is generated to see which of the techniques work the best;
3. model building- A learning model is built according to the results of the previous steps;
4. model assessment-Then, the model is trained and tested to assess how good it works

5. **Evaluation:**

Usually, by now, we ought to have one or two model that works well with the dataset, from a data scientist's perspective. It is imperative to evaluate the model, and to check if it conquers the business goals, review the steps executed to construct the model, before the final deploying of the model. A verdict about the handling of the data mining results should be reached.

Different steps are-

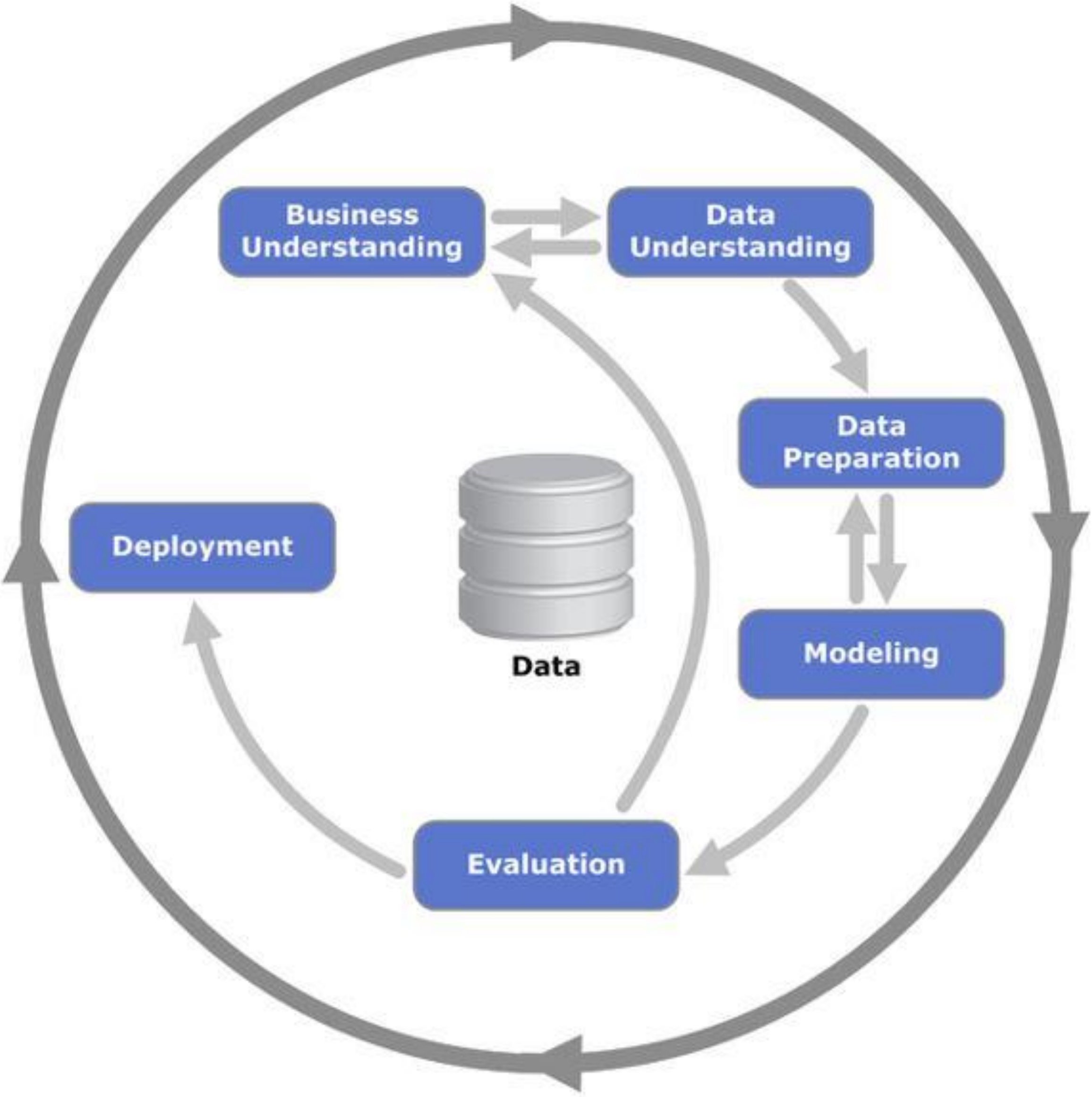
1. Results evaluation- Results are now checked;
2. process review- each process is reviewed;
3. next steps are determined- the further steps are determined.

6. **Deployment:**

The end of the project is not the construction of a model. The customer should be able to apply the knowledge acquired. The deployment phase is as easy as generating a report or as complex as implementing a repeatable data mining process. In any case, it is important to understand up front what actions will need to be carried out in order to actually make use of the created models.

Different steps are-

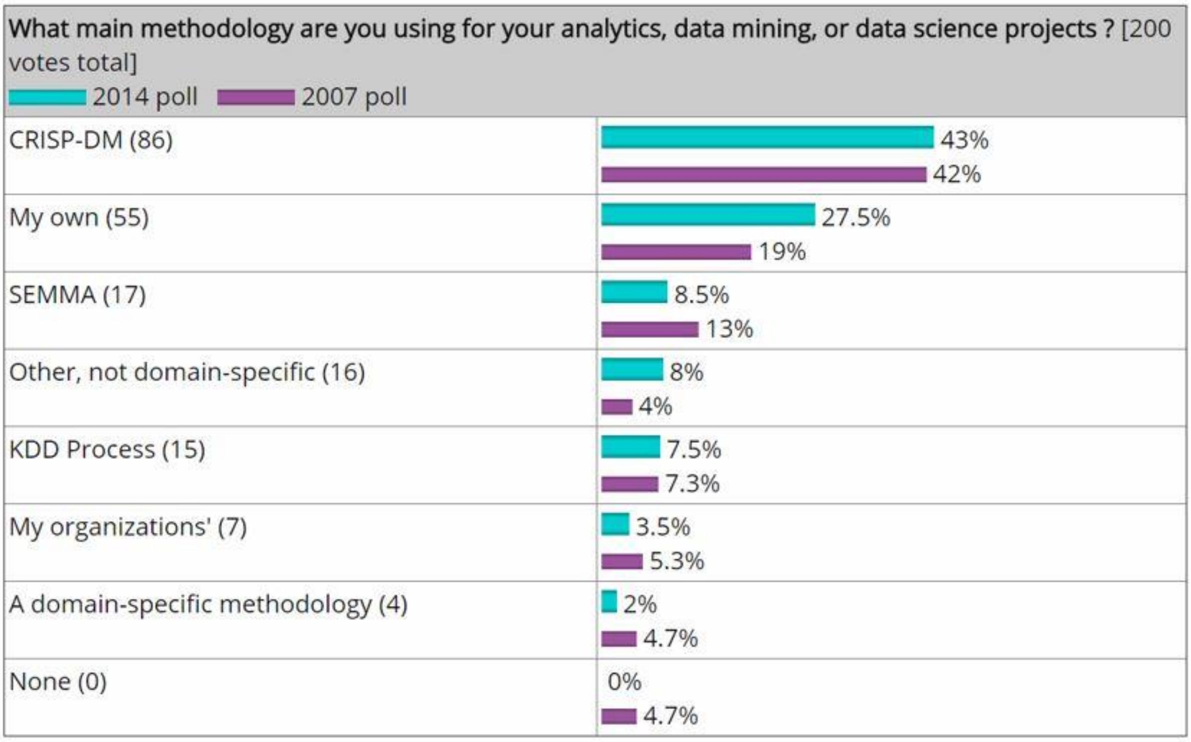
1. Deploying plan-this may involve beta testing;
2. monitor and maintain plan-check regularly to gauge the error/bias that may have been introduced during runtime;
3. final report production- showcasing the result;
4. project review- Reviewing the project as a whole and decide if any changes need to be made.



To recap, the six phases are:

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

(Piatesky, 2014)The most commonly used method for data mining and data science research, according to KDnuggets polls. Since its creation, It hasn't changed much despite its popularity.



What main methodology are you using for your analytics, data mining, or data science projects ? [200 votes total]
🟦 2014 poll 🟪 2007 poll

CRISP-DM (86)	43% 42%
My own (55)	27.5% 19%
SEMMA (17)	8.5% 13%
Other, not domain-specific (16)	8% 4%
KDD Process (15)	7.5% 7.3%
My organizations' (7)	3.5% 5.3%
A domain-specific methodology (4)	2% 4.7%
None (0)	0% 4.7%

3. Critique

CRISP-DM helps in maintaining work flow in data mining projects as it is work flow's standard description. Being the first step in the data science model, CRISP-DM has made a significant impact on data analysis by bringing order. However, CRISP-DM provides a approach which zeroes in on the task but there may be issues of team and communication.

Benefits

CRISP-DM is such a common sense and natural process that, when students were asked to do a data science project without project management directions, they "tended toward a CRISP-like methodology and identified the phases and did several iterations." Moreover, teams that were trained and explicitly told to implement CRISP-DM performed better than teams using other approaches (Saltz, Shamshurin, & Crowston, 2017).

William Vorhies (a CRISP-DM creator) argues that because all projects of data science begins only with the understanding of the business, have collect and comb through the data, and data science techniques needs to be applied(Vorhies, 2016).

CRISP-DM doesn’t need a lot of training, organizational role changes, or controversy for a successful implementation.

The literal first step, i.e. *Business Understanding,* makes sure that our technical jobs aligns with our business needs and data scientists stay back from going into a problem before all the business goals are understood. The project’s consideration transition to maintenance and operations; are assessed in its final step, i.e. *Deployment.*

There are many advantages of Agile. The end-user can gain an in-depth understanding of the data and the problems while going through various steps every time. The following cycles are fed the results of all previous cycles to gain a clear sight of the data.

Disadvantages

Contrary to Vorhies, since the CRISP-DM Model is a predecessor to Big Data, it is believed that an old model cannot support the needs of the modern world. However, Some people think that this particular model suffers from the same problem as the Waterfall model, in the fact that it is a hinderance to the fast iterations. Although that is the biggest problem, the one that everybody observes is that to maintain the sequential nature which heavily relies on documentation. Perhaps most significantly, CRISP-DM is not a true project management methodology because it implicitly assumes that its user is a single person or small, tight-knit team and ignores the teamwork coordination necessary for larger projects (Saltz, Shamshurin, & Connors, 2017).

Therefore, structure should be added to help coordinate teamwork.

4. Conclusions and Recommendations

CRISP-DM is a great starting framework for those who are looking to understand the general data science process. It likewise may serve individual and small teams well, and if augmented with other project management approaches, might suit larger teams. Specifically, emerging approaches that combine agile project management and CRISP-DM are likely to be more effective. (CRISP-DM Guide.)

1. INTRODUCTION

“Big Data is not about the data”. Gary King of Harvard University made this comment while making the point that the real value of collecting large volumes of data is in the extent to which interesting knowledge can be extracted from the data. While it is much touted that Big Data is the new gold, we believe that Big Data is rather the new gold ore. Data alone does not guarantee access to actionable insights. Rather, value is created when in-depth insights are derived from the data (gold ore) to enable one to generate meaningful knowledge (refined gold).

The critical care sector generates bountiful data around the clock, which can paradoxically complicate the quest for information, knowledge, and ‘wisdom’. The accumulation of clinical data has outpaced the capacity for effective aggregation and analysis aiming to support clinical quality, patient safety and integrated patient care. Intelligent data analysis promises a more efficient representation of the complex relations between symptoms, diseases and treatment. Additionally intelligent data analysis hopes for a reduction of cost of care and faster design and implementation of clinical guidelines. In this respect, the secondary use of clinical and operational data could support comparative effectiveness research, data mining, and predictive analytics. Commonly used data analysis platforms in clinical practice, frequently only provide support for data integration and monitoring, leaving all the analysis and decision taking to the clinical end-users. The clinical end-user is not in the position to constantly monitor and process the large amounts of data generated by patient monitoring and diagnostics. The potential of predictive analytics is to provide the clinical end-user with validated medical decision support and ultimately leading to more Predictive, Preventive and Personalized Medicine—PPPM. PPPM is an integrative concept in health care that enables to predict individual predisposition before onset of the disease, to provide targeted preventive measures and create treatment algorithms tailored to the person. PPPM relies on the potential of large amounts of heterogeneous data collected in medical environments (electronic health records, medical texts and images, laboratory tests etc), but also from external data of increasingly popular wearable devices, social media etc. Data driven predictive algorithms often fail to provide self explanatory models due to high-dimensionality and high-complexity of the data structure leading to unreliable models. Also, successful predictive analytics and application of cutting edge machine learning algorithms often demands substantial programming skills in different languages (e.g. Python or R). This migrates modeling from the domain expert to the data scientist, often missing the necessary domain expertise, and vice versa, domain experts are not able to perform ad hoc data analyses without the help of experienced analysts. This leads to slow development, adoption and exploitation of highly accurate predictive models, in particular in medical practice, where errors have significant consequences (for both patients and costs). In this paper, we address this problem by exploring the potential of visual, code free tools for predictive analytics. We also review the potential of visual platforms (RapidMiner, Knime and Weka) for big data analytics. As a showcase, we integrated the MIMIC-II database in the RapidMiner data analytics platform. Data extraction and preparation was performed on a Hadoop cluster, using RapidMiner’s Radoop extension. Further, we used RapidMiner Studio in order to develop several processes that allow automatic feature selection, parameter optimization and model evaluation. The process compared several learning methods (Decision Stump, Decision Tree, Naive Bayes, Logistic Regression, Random Forest, AdaBoost, Bagging, Stacking, Support Vector Machine) in association with feature weighting and selection quantitatively assessed in terms of Correlation, Gini Selection, Information Gain and ReliefF.

Data source

Data are available from the MIT Institutional Data Access / Ethics Committee for researchers who meet the criteria for access to confidential data. The MIMIC II Waveform Database (bedside monitor trends and waveforms) is freely available from http://physionet.org/mimic2/. The MIMIC II Clinical Database (all other elements of MIMIC II) is available under terms of a data use agreement (DUA). The entire analytical process as described in the paper is attached as XML.

Predictive algorithms

Predictive algorithms used are-

- **Naive Bayes (NB)**
- **Decision trees (DT**
- **Logistic regression (LR)**
- **Support Vector Machines (SVMs)**
- **Ensemble (meta-learning) methods**
- **Random Forest (RF)**

among others

2. KEY HIGHLIGHTS

With the accumulation of large amounts of health-related data, predictive analytics could stimulate the change of reactive medicine towards Predictive, Preventive and Personalized (PPPM) Medicine can be stimulated by prediction science with the collection of “big” data, ultimately affecting both cost and quality of care. However, easy translation of data-driven methods limits high-dimensionality and high-complexity of the data from being involved with clinically relevant models. Substantial programming skills are needed which limits its direct exploration by medical experts. This leaves a gap between potential and actual data usage. The authors addressed this problem by focusing on visual environments, easily accessible by the medical community. A framework was developed by the data from critical care patients by integrating the MIMIC-II database in RapidMiner supporting scalable predictive science using RapidMiner’s Radoop extension. The ETL process (Extract, Transform, Load) was initiated by retrieving data from the MIMIC-II tables of interest. As use case, correlation of platelet count and ICU survival was quantitatively assessed. Using Hadoop and RapidMiner, robust processes are constructed for automatic building, parameter optimization and evaluation of various predictive models, under different feature selection schemes. This environment is brilliant for scalable predictive analysis in health research because these processes can be easily adopted in other projects.

3. CONCEPTUAL FRAMEWORK

Some of the key steps are:-

Data extraction, pre-processing and exploratory analyses

Data is extracted, processed and explored in this step.

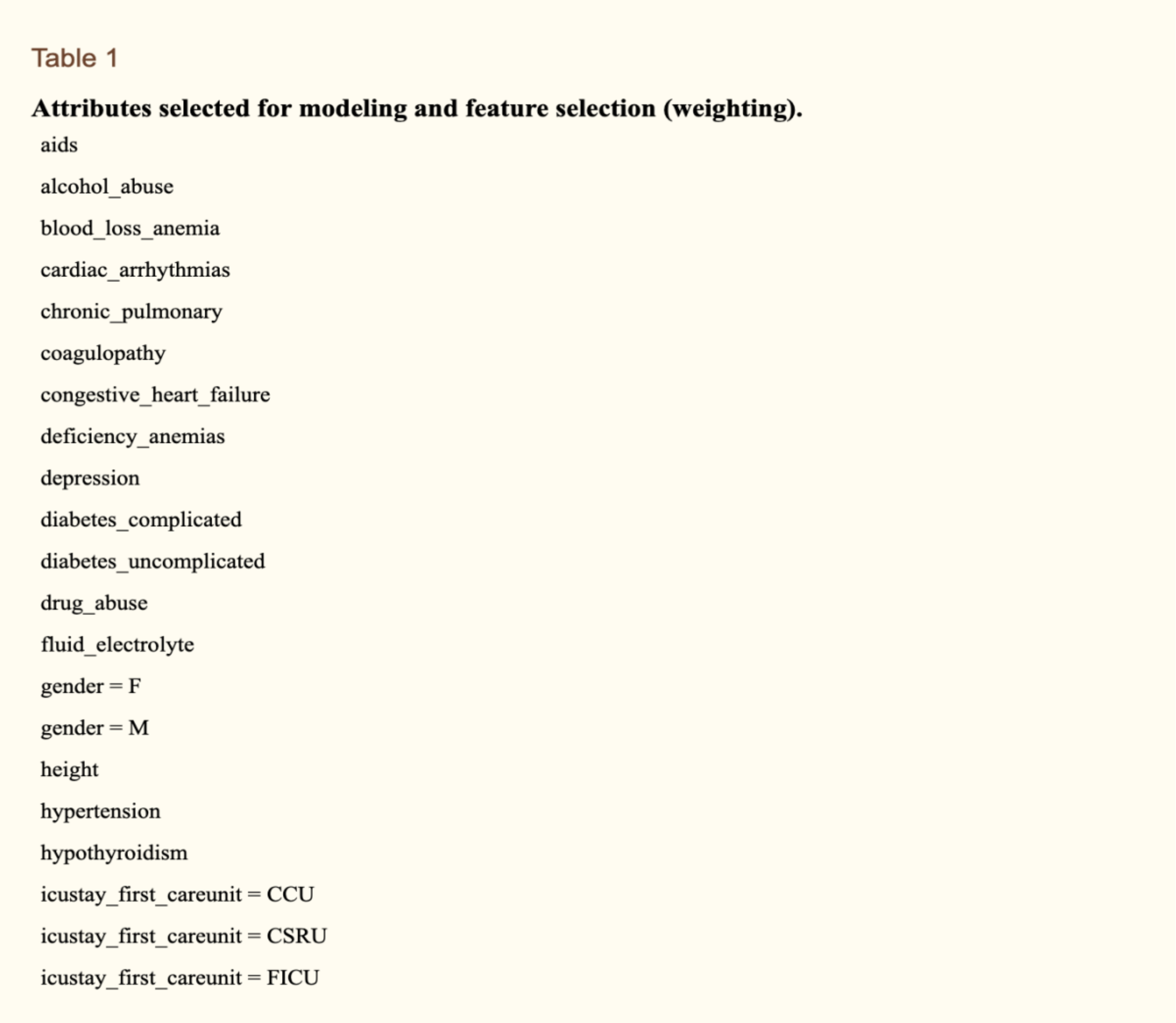


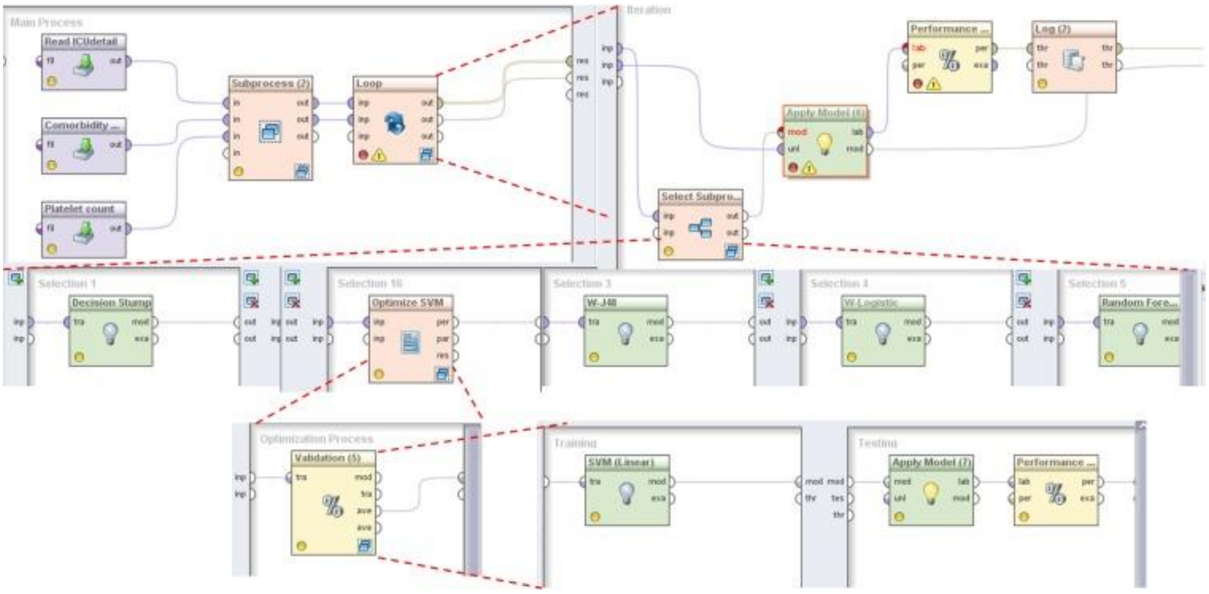
Table 1

Attributes selected for modeling and feature selection (weighting).

aids
alcohol_abuse
blood_loss_anemia
cardiac_arrhythmias
chronic_pulmonary
coagulopathy
congestive_heart_failure
deficiency_anemias
depression
diabetes_complicated
diabetes_uncomplicated
drug_abuse
fluid_electrolyte
gender = F
gender = M
height
hypertension
hypothyroidism
icustay_first_careunit = CCU
icustay_first_careunit = CSRU
icustay_first_careunit = FICU

Automatic model building, parameter optimization and evaluation

The model is built. Then, parameter optimized and each algorithm is evaluated to find the best performing algorithm.



Evaluation

The result from the best performing algorithm is checked and its error evaluated

Performance and Feature Selection

In this step the best performance value and the best features for machine learning selected

4. KEY PROCESSESS

In the main process, basic pre-processing was performed in a *Sub-process* operator after the data was taken out from the 3 data sets (Platelet count, Comorbidity, ICUdetail). A macro iterating over a user defined interval defines the final operator in the *Loop*. in a single process execution, looping was allowed to run, through multiple algorithms resulting in model building and evaluation in a single process execution. *Select Subprocess*, *Performance* and *Log* makes up the inner operators of the *Loop*. The *Apply model* took as an input the model provided from the *Select Sub-process* provides a model that is used as the input for the *Apply model* and, from the previous process level, holdout data that was forwarded

Area Under the Precision Recall Curve values were calculated when the model was applied. The RapidMiner-R extension was used allowing incorporation of R scripts within the Execute-R operator, based on the PRROC R package, because RapidMiner does not provide AUPRC calculation. The *Select Sub-process* operator, consisted of an inner operator structure that could be iteratively executed for user defined number of times to enable building and evaluation of multiple models in one process execution. 17 subprocesses were defined which contains one predictive algorithm. The Loop operator provides the execution order, which is taken care by the iteration macros.

Additionally, parameter optimization is allowed by this structure, if needed. Evolutionary algorithm is implemented for parameter optimization of SVMs

5. ADVANTAGES

The advantage is the representation of "real world", in which no protocol of study has been performed in collecting data, of using this type of database. Its strict inclusion and exclusion criteria is the reason for which many interventional trials have been criticized for. Clinical researchers rarely achieve the required expertise in SQL and the database is underemployed. A code-free UI and availability both in the cloud and as an open-source client/server platform is main advantages of The RapidMiner platform. RapidMiner has the depth adapted to the researchers' need to get info with a platform having more than 1,500 methods of all stages of the predictive science's life cycle. Time-to-insights reduced and guide for best practices for data analysts, analyzing the behaviour to make "wisdom of the crowds" makes RapidMiner optimizable. Based on 10 criteria, including completeness of focus and ability to execute data, SAS, IBM, KNIME, and RapidMiner lead in Gartner 2015 Magic Quadrant for Advanced Analytics Platforms.

Modeling with classification, evaluating predictive accuracy of models, visualizing the performance of models, and ranking patients by the predicted confidence for a clinical decision is enabled by data mining on ICU-patient data, in order to select the best candidates for a therapy.

6. CHALLENGES

Several limitations are there to be fixed.
Firstly, the study is retrospective with predicted drawbacks. For instance, there were some patients ignored which may be responsible for bias because the included cohort does not represent the whole study population. However, included and excluded patients are similar in many characteristics, making our patients representative of the target population.

Secondly, ICU patients were heterogeneous including medical, surgical and cardiac surgical patients the prognostic value of platelet count for survival will narrow down the study population requires further investigations.

Thirdly, other unknown factors may exist to confound the prognostic value of platelet count, even though every effort has been made to adjust for the confounding factors by using multivariate analysis. They made use of ICU mortality rather than the more commonly used ones such as 28-day and 90-day mortality as the study goal. This is because data may not be available after ICU discharge in the MIMIC-II database.

7. CONCLUSION

The integration of the MIMIC-II database through RapidMiner, which enables scalable predictive analysis of clinical information from ICU patients in a visual, code free surroundings. The proposed integration is seamless manipulation, data extraction, preprocessing and predictive analytics of huge amounts of data, using visual tools without the need for writing a single line of code which are the most import features. This tool can eventually bridge the gap between potential and actual usage of medical data. This approach helps in the development of data lakes (large storage repositories that keeps information in its native format until it is required) becoming attractive platforms in research facilities around the world. By defining several processes for automatic building of multiple models, parameter optimization, feature selection and model evaluation which is used as a showcase of the proposed environment we demonstrated a prognostic value of platelet count in critically ill patients,. These processes are strong enough to provide effective research in a variety of clinical research questions with little or no adoptions.

References

1. Van Poucke S, Zhang Z, Schmitz M, et al. Scalable Predictive Analysis in Critically Ill Patients Using a Visual Open Data Analysis Platform. *PLoS One*. 2016;11(1):e0145791. Published 2016 Jan 5. doi:10.1371/journal.pone.0145791
2. P. Kalgotra and R. Sharda, "Progression analysis of signals: Extending CRISP-DM to stream analytics," 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, 2016, pp. 2880-2885.
3. "https://pdfs.semanticscholar.org/9735/4f88e871fb01ee66aa0be8b71562c48bc181.pdf"
4. "https://www.the-modeling-agency.com/crisp-dm.pdf"
5. "https://scholarspace.manoa.hawaii.edu/bitstream/10125/41273/1/paper0124.pdf"
6. "https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html"
7. "https://scholarspace.manoa.hawaii.edu/bitstream/10125/41273/1/paper0124.pdf"
8. "https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome"