Overview: Data Mining Pipeline

Proseminar Data Mining

Marija Juodyte
Faculty of Informatics
Technical University of Munich
Email: marijajuo@gmail.com

***Abstract*— This paper covers the general applications of data mining and the abstract tasks coming along the way. CRISP-DM and Fayyad's Pipeline models are being analyzed and each of the pipeline steps are described in detail.**

***Index Terms*— Data Mining, Pipeline, Fayyad Model, CRISP-DM, Knowledge Discovery in Databases, Business understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment**

I. Introduction

The exponential growth of data has lead to an obvious improvement in people's lives. However, this leads us to the question on how to determine which information is essential and how to efficiently extract it. The bigger the data set, the more important it is to pick out a good strategy to process the data.

There has been some research concerning this area. As the computers advanced, this has enabled to look at the data and statistics from a whole new perspective. In 1989 Gregory Piatetsky-Shapiro organizes and chairs the first Knowledge Discovery in Databases (KDD) workshop. In 1995, it became the annual ACM Special Interest Group on Knowledge Discovery and Data Mining Conference (SIGKDD). Following that, "From Data-Mining to Knowledge Discovery in Databases" [1] paper is being released in 1996 by U. Fayyad, G. Piatetsky-Shapiro, P. Smyth formulating the first concrete steps in a data mining project. KDD is a non trivial multi-step process to extract information from the given data. It's a result of various research fields from Informatics and Mathematics like Databases, Machine Learning, Statistic, Data visualizing and High Performance Computing. KDD entails all the steps from where the data is being picked up to actual processing to extract the information. Data Mining is only a step in a KDD process, but often used as a synonym.

KDD has to deal with lots of challenges, like working on gigantic databases with complex relationships between attributes. Quite often they have to cover cases of missing data, fake data and noise, overfitting and underfitting.

In the year 2000 a paper on "Cross Industry Standard Process for Data Mining" (CRISP-DM) [2] is being released which defines the general tasks in an industrial data mining project. The steps of a data mining process have remained mainly the same until these days.

KDD can be applied in various fields: patient diagnosis in medicine, DNA and Protein Sequentializing in biology, foreseeing the behaviour of different user groups in marketing, identifying possible fraud occurrences and diagnosing, and foreseeing product defects.

This paper shall give an overview on the steps of a data mining project. The common tasks of a data mining project combine different methods to form a pipeline. We shall go through all of the steps illustrated by an example of a large clinical database - specifically, data accumulated on 3902 obstetrical patients evaluated for factors potentially contributing to preterm birth. [3]

II. methods

Generically speaking, a pipeline has inputs go through a number of processing steps chained together in some way to produce some sort of an output. In the pipeline, each step logically follows the next step providing an outcome.
The KDD-Model pipeline consist of the following steps [1]:

1) application domain understanding,
2) data selection,
3) data cleaning and preparation,
4) data reduction and transformation,
5) suitable data mining methods selection,
6) data mining step,
7) interpretation of the found results,
8) deployment of the found knowledge.

Figure 1 represents the CRISP-DM pipeline steps:

1) business understanding,
2) data understanding,
3) data preparation,
4) modeling,
5) evaluation,
6) deployment.

CRISP-DM model has merged the steps 3 and 4 of KDD to one data preparation step and steps 5 and 6 to one modeling step.

We are going to dive into each step of CRISP-DM and illustrate it with an example on a data mining project at the Duke University Medical Center. [3]

A. Business Understanding

This initial phase focuses on understanding the project and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives. [2] When determining business objectives the main goal is to understand

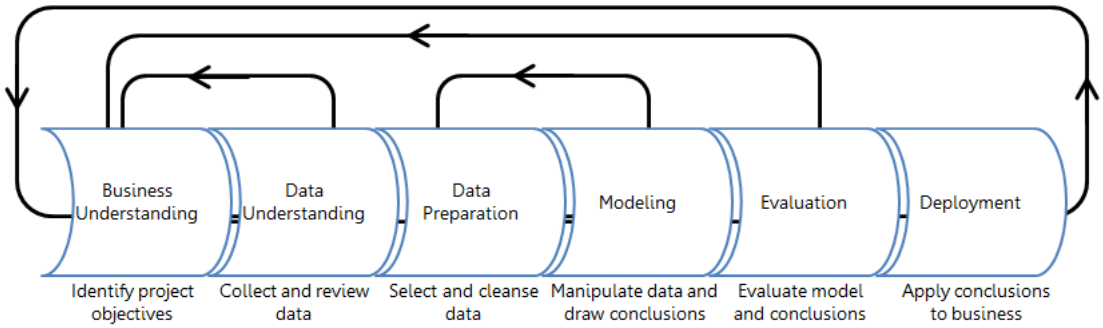

Fig. 1. Cross Industry Standard Process for Data Mining (CRISP-DM) [2]

what the customer actually wants: business success criteria. This consists of determining what the customers and what their interests are and what information are they trying to gain from the data mining project. In addition to that, one needs to consider the inventory of resources and determining the requirements, assumptions, constraints, risks, costs and benefits. In the end one needs to define data mining goals and data mining success. In general this step is an initial assessment of tools and techniques. If the step is being neglected, a likely consequence would be wasting a lot of time answering the wrong questions.

In our example the examined data set is clinical data about the patients experiencing preterm birth. Prematurity remains the most common cause of low birthweight and associated morbidity and mortality. The eventual goal of this knowledge discovery effort is to identify factors that will improve the quality and cost effectiveness of perinatal care. [3]

B. Data Understanding

The data understanding phase starts with initial data collection and proceeds with activities that enable one to become familiar with the data, identify data quality problems, discover first insights into the data, and/or detect interesting subsets to form hypotheses regarding hidden information. [2] In this step one needs to collect the Initial Data, describe the data, explore the data and verify the data quality. The things to be considered is if the data type is static, nominal or ordinal. In addition to that what data structures does it take: object, set, dynamic, stream-data or time queues. Another important aspect is considering what the data source actually is and how to extract the data from it. The data can be stored in databases, data warehouses or world wide web. In Order to create a data warehouse one needs to understand the size of the data sets as well. This means that one needs to consider the number of data sentences, number of attributes, number of allowed attribute values. Quite often one comes across poor inconsistent, not exact, mixed data. One then also needs to decide what steps will need to be taken in the data preparation phase. [4] Data Visualizing is one thing helping along the way.

In our example database identified for mining was the computer-based patient record system known as The Medical Record, or TMR. TMR is a comprehensive longitudinal Computerized Patient Record System developed at Duke University over the last 25 years. The data collected in TMR include demographics, study results, problems, therapies, allergies, subjective and physical findings, and encounter summaries. TMR's data structure uses a proprietary class-oriented approach which stores all of the patient's information in a single record. This database continues to serve as the repository for a regional perinatal computerized patient record that is used in inpatient and outpatient settings [5]. The on-line Duke perinatal database contains comprehensive data on over 45,000 unique patients collected over nearly 10 years. Additional patient data from the previous decade is also available on tape archive. This computerized repository contains more than 4,000 clinical variables collected on over 20,000 pregnancies and births from a five county area, making it one of the largest and most comprehensive obstetrical datasets available for analysis in the United States. [3]

C. Data Preparation

The data preparation phase covers all activities needed to construct the final dataset [data that will be fed into the modeling tool(s)] from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools. [2]

1) Data Selection: In the data selection phase it is important to determine why is a specific data included or excluded. The selection is being applied to multiple levels: one needs to decide which entries should be selected and which attributes of the entries are relevant. Adding irrelevant or distracting attributes to a dataset often confuses machine learning systems. [6]

In the Duke University Database the data warehouse was created on a centralized server dedicated to fielding data mining queries. [3] The clinical data was mapped from the data base to a personal computer. The warehouse contained 45,922 patient records. These records included 215,626 encounters; 1,757,118 historical data elements; 3,898,887 individual lab results; 217,453 problems and procedures; and 3,016,313 subjective and physical findings. A sample of two-year dataset

(1993-1994) from the data warehouse was then created to be mined for knowledge discovery.

2) Data Cleaning: As each variable is added to the dataset, it is cleaned of erroneous values, data inconsistencies, and formatting discrepancies. The crucial part of data cleaning is usually discretization of continuous data and/or grouping them out to permit statistical analysis. Additionally, it is important to ensure that multiple values for the same variable are not present and that the entries with missing or extreme values need to be taken into consideration. Normalizing the data in order to fit the standards should also be a key point.

In our example the cleaning process is accomplished using Paradox Application Language scripts [3] to selectively identify problems and correct the errors. If multiple values for the same variable existed, the value that was recorded closest to delivery or conception, depending on perceived data quality for the particular variable, was loaded into the final dataset. The data set has also been scanned converting alphanumeric fields into numerical variables. A final script identified missing values and prompted the user to either substitute them with an average value for the variable, or to delete the subject from the dataset. The test dataset extracted from the data warehouse contained data regarding 3902 births occurring between January 1, 1993, and December 31, 1994. The data cleaning programs used in creating the dataset revealed that 8.74% of the total values in the database were unusable for the purposes of the factor analysis (Table I). Free text stored in place of a coded data phrase, missing values, incomplete dates accounted for the largest amount of unusable data. Other causes why the data has been unused included out of range values such as invalid heights, format discrepancies and data inconsistencies such as two very different heights for one patient, combined in one group.

D. Modeling

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, going back to the data preparation phase is often necessary. [2]

Modeling is the step where the real data mining algorithms are applied. In this phase a correct selection of an algorithm is needed by considering the required input and output. Data mining involves six common classes of tasks: [1]

1) Anomaly detection: (outlier/change/deviation detection) The identification of unusual data records, that might be interesting or data errors that require further investigation. This Topic has already been covered in the Data Cleaning subsection.

2) Association rule learning: (dependency modeling) searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Supermarkets may determine which products are frequently bought together using association rule learning. The gained knowledge can then be used for the marketing purposes (market basket analysis). Following the original definition by Agrawal, Imieliski, Swami [7] the problem of association rule mining is defined as:

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of n binary attributes called items. Let $D = \{t_1, t_2, \ldots, t_m\}$ be a set of transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of the items in I. A rule is defined as an implication of the form:

$X \Rightarrow Y$, where $X, Y \subseteq I$

Every rule is composed by two different sets of items, also known as itemsets, X and Y, where X is called antecedent or left-hand-side (LHS) and Y consequent or right-hand-side (RHS).

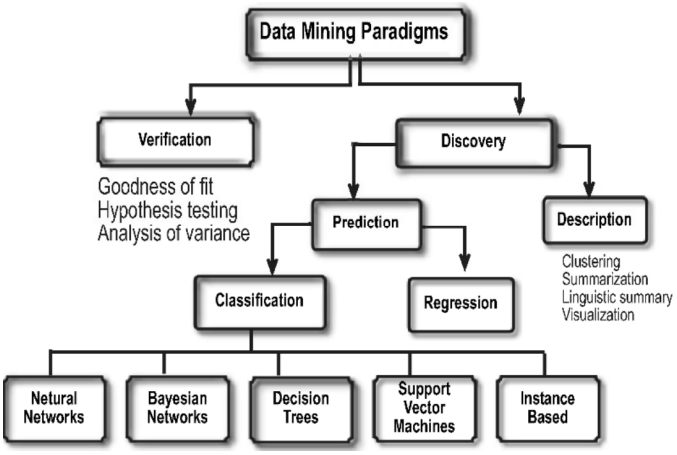

Fig. 2. Data Mining Paradigms [10]

3) Clustering: is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data. Clustering belongs to the discovery and description paradigms of data mining (Figure 2). The general idea behind clustering is to group the most similar entities in the data matrix together. One can choose one of the two ways to do so: either compute the similarities (i.e. correlation coefficient) or the dissimilarities (i.e. distance measures).

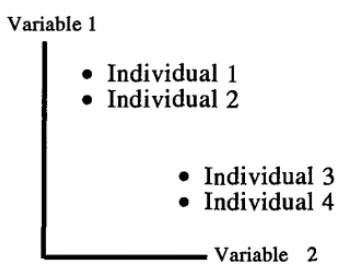

Fig. 3. Illustration of Distance and Similarity Among Four Individuals in a Two-Variable Property Space [8]

Figure 3 shows the position of four individuals in a two-variable property space composed of variable 1 and variable 2.

TABLE I

CHARACTERISTICS OF THE PRODUCTION SYSTEM DATABASE AND THE CLINICAL DATA WAREHOUSE [3]

Reason Unusable	**Example**	**Count**	**% of Total Values**
Missing values when required	Ward clerk did not enter or data item was not collected 2,213 2.95%	2,213	2.95%
Incomplete dates	Dates preventing calculations, e.g. ??/??/94	249	0.33%
Free-text in place of a coded data phrase	Ward clerk enters free text for an item in place of code from the data dictionary	4,071	5.43%
Other errors	Out ofrange values, format discrepancies, data inconsistencies	16	0.02%
	Totals:	6,549	8.74%

It is clear from visual inspection that individuals 1 and 2 would be grouped into cluster 1, and individuals 3 and 4 into cluster 2. In terms of similarity or dissimilarity, individuals 1 and 2 have high correlations (low distance) between each other, as do individuals 3 and 4. Comparing individuals 2 and 3, however, we see that they are quite dissimilar (exhibiting higher distance between them) on both variable 1 and variable 2, showing that their correlation is low. Thus, individuals close together in the property space show low distance (high correlation) on all dimensions, while those far apart show high distance (low correlation).' This means that distance and correlation measures are the inverse of one another. [8] One way how to calculate the distance is:

$D^2 = \sum_{i=1}^{K} (X_{A_i} - X_{B_i})^2$

Here D is the distance, K is the number of variables (dimensions), X_{A_i} signifies the value of variable i for object A, and X_{B_i} signifies the value of variable i for object B. [8]

4) Classification: is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

Examples of classification algorithms include linear classifiers, Fisher's linear discriminant, Logistic regression, Naive Bayes classifier, Perceptron, Support vector machines, Least squares support vector machines, Quadratic classifiers, Kernel estimation, k-nearest neighbor, Boosting (meta-algorithm), Decision trees, Random forests, Neural networks and Learning vector quantization.

A large number of algorithms for classification can be phrased in terms of a linear function that assigns a score to each possible category k by combining the feature vector of an instance with a vector of weights, using a dot product [9]. The predicted category is the one with the highest score. This type of score function is known as a linear predictor function and has the following general form:

$\text{score}(\mathbf{X}_i, k) = \boldsymbol{\beta}_k \cdot \mathbf{X}_i$

5) Regression: Regression algorithm attempts to find a function which models the data with the least error that is,for estimating the relationships among data or datasets. It finds best matching curves to data points e.g. simple, multiple linear, non-linear regression, logistic regression. An important part of regression is finding a suitable interpolating function. Regression belongs to the discovery and prediction data mining paradigms. (Figure 2)

6) Summarization: providing a more compact representation of the data set, including visualization and report generation.

In the example of the Duke University [3], exploratory factor analysis for data mining has been chosen because it had previously been used successfully to explore claims and financial databases in the field. [11] Factor analysis has also been chosen which identifies which data elements can be combined to explain variations between patient groups. This mining technique is appropriate in research problems in which a large number of subjects are compared on a set of variables for which there is no designation of independence or dependence [12].

E. Evaluation

At this stage in the project, one has built a model that appears to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to thoroughly evaluate it and review the steps executed to create it, to be certain the model properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached. [2]

Data mining can unintentionally be misused, and can then produce results which appear to be significant; but which do not actually predict future behaviour and cannot be reproduced on a new sample of data and bear little use. [2] Investigating too many hypotheses and not performing proper statistical testing can lead to overfitting. [13]

To overcome this, the evaluation uses a test set of data on which the data mining algorithm was not trained. The learned patterns are applied to this test set, and the resulting output is compared to the desired output. For example, a data mining algorithm trying to distinguish "spam" from "legitimate" emails would be trained on a training set of sample e-mails. Once trained, the learned patterns would be applied to the test set of e-mails on which it had not been trained. The accuracy of the patterns can then be measured from how many e-mails they correctly classify. A number of statistical methods may be used to evaluate the algorithm, such as ROC curves.

If the learned patterns do not meet the desired standards, subsequently it is necessary to re-evaluate and change the preprocessing and data mining steps. If the learned patterns do meet the desired standards, then the final step is to interpret the learned patterns and turn them into knowledge.

In the Duke University example [3] Exploratory Factor

Analysis successfully extracted dataset from the data warehouse. All analyses used listwise deletion of cases with missing values, principal components analysis, and varimax rotation. Preliminary results identified three latent factors that accounted for 48.9% of the variance in the sample. Table **??** shows the results of the factor analysis along with the variables contributing the greatest weight to the factor in the sample dataset. [3]

F. Deployment

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. It often involves applying live models within an organizations decision making processes for example, real-time personalization of Web pages or repeated scoring of marketing databases. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. In many cases, it is the customer, not the data analyst, who carries out the deployment steps. However, even if the analyst will carry out the deployment effort, it is important for the customer to understand up front what actions need to be carried out in order to actually make use of the created models. [2]

Knowledge gained will need to be organized and presented in a way that the customer can use it. Applying live models within an organizations decision making processesfor example, real-time personalization of Web pages or repeated scoring of marketing databases. Can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

In Duke University example [3] newly discovered relationships found in clinical databases such as these will potentially lead to better understanding between observations and outcomes in perinatal care and other fields of medicine. By implementing CPRS data warehousing, new medical hypotheses can be generated for predicting and preventing preterm birth and other adverse health outcomes.

III. RESULTS

Each of the steps in the pipeline have produced some results which were later used in the following steps of the pipeline or have been deployed.

The data mining pipeline is iterative, therefore one can come back to the previous steps of the pipeline and repeat them at any point in time.

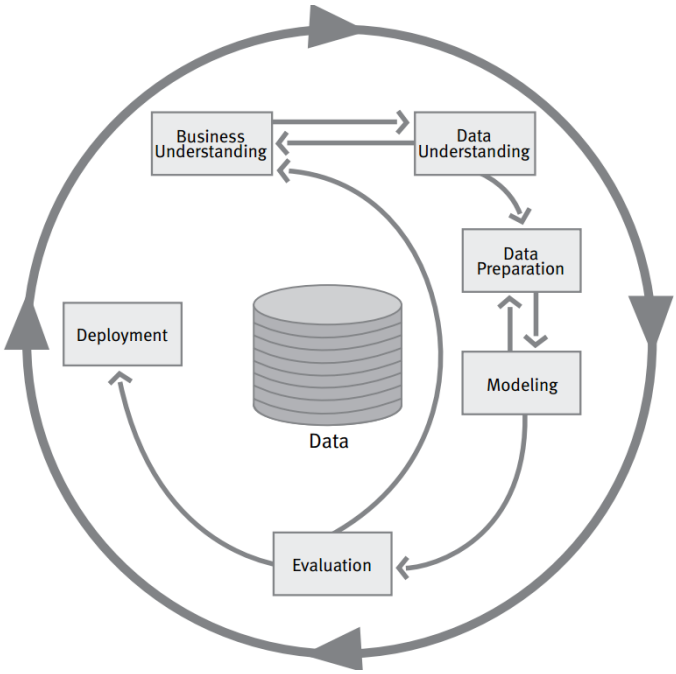

Fig. 4. Iterative Model of CRISP-DM [2]

The first step of the pipeline - Business understanding outputs the business objectives, business success criteria inventory of resources, requirements, assumptions and constraints as well as the most important things - data mining goals.

The Data Understanding step of the project outputs initial data collection report, data description report, data exploration report as well as data quality report.

Data Preparation step offers dataset description, rationale for inclusion/exclusion, data cleaning report, generated records and merged, reformatted data.

Modeling offers modeling assumptions, test design, model description and revised parameter settings.

Evaluation step offers approved models, review of process and final decision on what to do with the extracted knowledge.

After the deployment step one should have a deployment plan, monitoring and maintenance plan, ways to present the gain of knowledge to the business side/customers, and write good documentation on the experience.

The final result of the pipeline is usually a data product or a set of decisions and their supports. As shown in figure 4, in case of incorrect results, going back to the previous steps is possible at any step of a project and the whole model is iterative.

IV. DISCUSSION

Different steps of the pipeline take different amounts of effort. One can see that the biggest part of the project takes the preparation of data, making around 60% of the relative effort. Due to erroneous or missing data the preparation is such a crucial part of the data mining project. The second place takes the understanding of the data. The Data mining process itself, evaluation of results, deployment of results take relatively similar amount of effort.

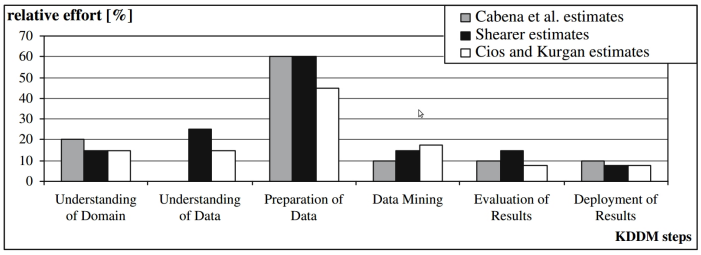

Fig. 5. KDDM relative effort [1]

V. Conclusion

- An important thing before building the pipeline is to explore the data before making any assumptions and running checks on it afterwards.
- The challenges of data mining algorithms are gigantic databases, complex relationships between attributes, missing data, fake data and noise, understanding the won patterns and integration in other Systems.
- Different cases require different data mining algorithms.
- Evaluation of the Data Mining results is crucial.
- A repetition and going back to previous steps may occur if incorrect results have been produced.

References

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data-mining to knowledge discovery in databases," *AI Magazine*, vol. 1, no. 3, 1996.
[2] P. C. (NCR), J. C. (SPSS), R. K. (NCR), T. K. (SPSS), T. R. (Daimler-Chrysler), C. S. (SPSS), and R. W. (DaimlerChrysler), "Cross industry standart process for data mining," 2000.
[3] D. F. L. M. P. M. L. K. G. R. P. J. W. H. P. M. L. H. M. Jonathan C. Prather, M.S. and P. W. Edward Hammond, "Medical data mining: Knowledge discovery in a clinical data warehouse," *Proc AMIA Annu Fall Symp*, pp. 101–105, 1997.
[4] H. D. E Rahm, "Data cleaning: Problems and current approaches," 2000.
[5] M. Burkett, "The tertiary center and health departments in cooperation: The duke university experience," *The Journal of perinatal and neonatal nursing*, vol. 2, pp. 11–19, 1989.
[6] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2005.
[7] T. S. A. Agrawal, R.; Imieliski, *Mining association rules between sets of items in large databases*, 1993.
[8] K. D. Bailey, "Typologies and taxonomies: An introduction to classification techniques," 1994.
[9] D. Mladeni, J. Brank, M. Grobelnik, and N. Milic-Frayling, "Feature selection using linear classifier weights: interaction with classification models," *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004.
[10] O. Maimon and L. Rokach, *Data-Mining and Knowledge Discovery Handbook*, 2005.
[11] L. Woolery and J. Grzymala-Busse, "Machine learning and preterm birth risk assessment," *Journal of the American Medical Informatics Association.*, vol. 1(6), pp. 439–446, 1994.
[12] D.-S. B and T. RG., "Basic and clinical biostatistics," pp. 227–228, 1994.
[13] D. M. Hawkins, "The problem of overfitting," *Journal of chemical information and computer sciences*, vol. 44, no. 1, pp. 1–12, 2004.