SNS COLLEGE OF TECHNOLOGY

Coimbatore-35.
An Autonomous Institution

COURSE NAME : DATA ANALYTICS

II YEAR/ IV SEMESTER

UNIT – I INTRODUCTION

Topic: *KDD*

Dr.K.Sangeetha
HoD
Department of Computer Science and Engineering

Methodologies for Data Analytics

- Methodology from Academia, KDD
- Methodology from industry, CRISP-DM (**CR**oss **I**ndustry **S**tandard **P**rocess.)

The KDD process proposes a sequence of nine steps.
In spite of the sequence, the KDD process considers the possibility of going back to any previous step in order to redo some part of the process. The nine steps are:

Methodology from Academia, KDD

1) Learning the application domain*: What is expected in terms of the application* domain? What are the characteristics of the problem; its specificities? A good understanding of the application domain is required.

2) *Creating a target dataset**: What data are needed for the problem? Which* attributes? How will they be collected and put in the desired format (say, a tabular data set)? Once the application domain is known, the data analyst team should be able to identify the data necessary to accomplish the project.

3) *Data cleaning and pre-processing**: How should missing values and/or outliers* such as extreme values be handled? What data type should we choose for each attribute? It is necessary to put the data in a specific format, such
as a tabular format

Methodology from Academia, KDD

4)Data reduction and projection*: Which features should we include to represent* the data? From the available features, which ones should be discarded? Should further information be added, such as adding the day of the week to a timestamp? This can be useful in some tasks. Irrelevant attributes should be removed.

5) ***Choosing the data mining function****: Which type of methods should be used?* Four types of method are: summarization, clustering, classification and regression. The first two are from the branch of descriptive analytics while
the latter two are from predictive analytics.

6) ***Choosing the data mining algorithm(s****): Given the characteristics of the problem* and the characteristics of the data, which methods should be used? It isexpected that specific algorithms will be selected.

7) Data mining*: Given the characteristics of the problem, the characteristics of* the data, and the applicable method type, which specific methods should be used?Which values should be assigned to the hyper-parameters ?The choice of method depends on many different factors: interpretability, ability to handle missing values, capacity to deal with outliers, computational efficiency, among others.

8) ***Interpretation****: What is the meaning of the results? What is the utility for* the final user? To select the useful results and to evaluate them in terms of the application domain is the goal of this step. It is common to go back to a previous step when the results are not as good as expected.

9) ***Using discovered knowledge****: How can we apply the new knowledge in practice?* How is it integrated in everyday life?This implies the integration of the new knowledge into the operational system or in the reporting system.

The CRISP-DM Methodology

1) ***Business understanding:*** *This involves understanding the business domain, being* able to define the problem from the business domain perspective, and finally being able to translate such business problems into a data analytics
problem.

2) ***Data understanding****: This involves collection of the necessary data and their* initial visualization/summarization in order to obtain the first insights, particularly but not exclusively, about data quality problems such as missing
data or outliers.

3) ***Data preparation****: This involves preparing the data set for the modeling tool*, and includes data transformation, feature construction, outlier removal, missing data fulfillment and incomplete instances removal.

The CRISP-DM Methodology

4)Modeling: *Typically there are several methods that can be used to solve* the same problem in analytics, often with specific data requirements. This implies that there may be a need for additional data preparation tasks that are method specific. In such case it is necessary to go back to the previous step. The modeling phase also includes tuning the hyper-parameters for
each of the chosen method(s).

5) ***Evaluation:*** *Solving the problem from the data analytics point of view is* not the end of the process. It is now necessary to understand how its use is meaningful from the business perspective; in other words, that the obtained solution answers to the business requirements.

6) ***Deployment:*** *The integration of the data analytics solution in the business* process is the main purpose of this phase. Typically, it implies the integration of the obtained solution into a decision-support tool, website maintenance
process, reporting process or elsewhere.

References

- Joao Moreira, Andre Carvalho, Tomás Horvath – "A General Introduction to Data Analytics" Wiley -2018
- Dean J, ―Big Data, Data Mining and Machine learning, Wiley publications, 2014.
- Provost F and Fawcett T, ―Data Science for Business, O'Reilly Media Inc, 2013.
- Janert PK, ―Data Analysis with Open Source Tools, O'Reilly Media Inc, 2011.
- Weiss SM, Indurkhya N and Zhang T, ―Fundamentals of Predictive Text Mining, Springer-Verlag London Limited, 2010.
- Marz N and Warren J,- Big Data, Manning Publications,2015
- Runkler T A, - Data Analytics: Models and Algorithms for Intelligent data analysis, Springer, 2012

https://www.edureka.co/blog/data-science-vs-big-data-vs-data-analytics/

Thank You!