**ALFRED NGAHU MUTHUKU**

**SCM225-0253/2024 BSc. OPERATIONS RESEARCH**

**INTRODUCTION TO DATA SCIENCE**

**ASSIGNMENT**

METHODOLOGIES

1. CRISP-DM

CRISP-DM stands for cross-industry process for data mining. The CRISP-DM methodology provides a structured approach to planning a data mining project. It is a robust and well-proven methodology. Its phases are as follows:

(a)Business Understanding:

The Business Understanding phase focuses on understanding the objectives and requirements of the project. Other tasks in this phase are foundational project management activities that are universal to most projects:

1. **Determine business objectives:** You should first “thoroughly understand, from a business perspective, what the customer wants to accomplish and then define business success criteria.
2. **Assess situation:** Determine resources availability, project requirements, assess risks and contingencies, and conduct a cost-benefit analysis.
3. **Determine data mining goals:** In addition to defining the business objectives, you should also define what success looks like from a technical data mining perspective.
4. **Produce project plan:** Select technologies and tools and define detailed plans for each project phase.

(b) Data Understanding:

Next is the Data Understanding phase. Adding to the foundation of *Business Understanding*, it drives the focus to identify, collect, and analyze the data sets that can help you accomplish the project goals. This phase also has four tasks:

1. **Collect initial data:** Acquire the necessary data and if necessary, load it into your analysis tool.
2. **Describe data:** Examine the data and document its surface properties like data format, number of records, or field identities.
3. **Explore data:** Dig deeper into the data. Query it, visualize it, and identify relationships among the data.
4. **Verify data quality:** How clean/dirty is the data? Document any quality issues.

**(c) Data Preparation**

is phase, which is often referred to as “data munging”, prepares the final data sets for modeling. It has five tasks:

1. **Select data:** Determine which data sets will be used and document reasons for inclusion or exclusion.
2. **Clean data:** Often this is the lengthiest task. Without it, you’ll likely fall victim to garbage-in, garbage-out. A common practice during this task is to correct, impute, or remove erroneous values.
3. **Construct data:** Derive new attributes that will be helpful. For example, derive someone’s body mass index from height and weight fields.
4. **Integrate data:** Create new data sets by combining data from multiple sources.
5. **Format data:** Re-format data as necessary. For example, you might convert string values that store numbers to numeric values so that you can perform mathematical operations.

**(d) Modeling:**

Here you’ll likely build and assess various models based on several different modeling techniques. This phase has four tasks:

1. **Select modeling techniques:** Determine which algorithms to try e.g. regression, neural net.
2. **Generate test design:** Pending your modeling approach, you might need to split the data into training, test, and validation sets.
3. **Build model:** As glamorous as this might sound, this might just be executing a few lines of code like “reg = Linear Regression (). fit (X, y)”.
4. **Assess model:** Generally, multiple models are competing against each other, and the data scientist needs to interpret the model results based on domain knowledge, the pre-defined success criteria, and the test design.

**(e) Evaluation:**

Whereas the Assess Model task of the Modeling phase focuses on technical model assessment, the Evaluation phase looks more broadly at which model best meets the business and what to do next. This phase has three tasks:

1. **Evaluate results:** Do the models meet the business success criteria? Which ones should we approve for the business?
2. **Review process:** Review the work accomplished. Was anything overlooked? Were all steps properly executed? Summarize findings and correct anything if needed.
3. **Determine next steps:** Based on the previous three tasks, determine whether to proceed to deployment, iterate further, or initiate new projects.

**(f) Deployment:**

A model is not particularly useful unless the customer can access its results. The complexity of this phase varies widely. This final phase has four tasks:

1. **Plan deployment:** Develop and document a plan for deploying the model.
2. **Plan monitoring and maintenance:** Develop a thorough monitoring and maintenance plan to avoid issues during the operational phase or post-project phase of a model.
3. **Produce final report:** The project team documents a summary of the project which might include a final presentation of data mining results.
4. **Review project:** Conduct a project retrospective about what went well, what could have been better, and how to improve in the future.

Your organization’s work might not end there. As a project framework, CRISP-DM does not outline what to do after the project also known as “operations”. But if the model is going to production, be sure you maintain the model in production. Constant monitoring and occasional model tuning are often required.

2. AGILE

The Agile methodology is a project management approach that involves breaking the project into phases and emphasizes continuous collaboration and improvement. Teams follow a cycle of planning, executing, and evaluating. Agile is technically not a methodology by itself, but rather a mindset for approaching how projects get done. It's not considered a methodology because Agile doesn't specify which tools and processes should be used.

Agile is, however, the umbrella term for many types of management methodologies. Scrum, Kanban, and Extreme Programming (XP) are each considered different Agile methodologies

3. LEAN

Lean methodology is a systematic approach to continuous process improvement through the elimination of waste such as defects, overproduction and waiting, etc. It originated in the manufacturing industry, specifically within Toyota, in the mid-20th century as part of the Toyota Production System (TPS), which sought to maximize value for the customer through the optimal use of resources.

Lean methodology facilitates an ongoing process of incremental adjustment, significantly accelerating product delivery by optimizing resources and effort and allowing teams to work efficiently and effectively.

It has 5 principles that are:

(a)Identify value

(b)Map the value stream

(c)Create a flow

(d)Establish pull

(e)Seek Perfection