COMPARISON OF METHODOLOGICAL APPROACHES: CRISP-DM VERSUS OSEMN METHODOLOGY USING LINEAR REGRESSION AND STATISTICAL ANALYSIS

A THESIS SUBMITTED TO
THE FACULTY OF ARCHITECTURE AND ENGINEERING
OF
EPOKA UNIVERSITY

BY

KETJONA SHAMETI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

JUNE, 2024

Approval sheet of the Thesis

This is to certify that we have read this thesis entitled **"Comparison of Methodological Approaches: CRISP-DM vs OSEMN Methodology using Linear Regression and Statistical Analysis"** and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

_____

Assoc. Prof. Dr. Arban Uka
Head of Department
Date: June, 26, 2024

Examining Committee Members:

Assoc. Prof. Dr. Dimitrios Karras (Computer Engineering) _____

Prof. Dr. Betim Çiço (Computer Engineering) _____

Dr. Florenc Skuka (Computer Engineering) _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name Surname: Ketjona Shameti

Signature: _____

ABSTRACT

COMPARISON OF METHODOLOGICAL APPROACHES: CRISP-DM VERSUS OSEMN METHODOLOGY USING LINEAR REGRESSION AND STATISTICAL ANALYSIS

Shameti, Ketjona
M.Sc., Department of Computer Engineering
Supervisor: Prof. Dr. Betim Çiço

AI has contributed in changing many industries, providing new and inventive solutions to complicated challenges. Nevertheless, efficient application of AI projects needs a structured and combinative technique in order to be updated with the latest advances in the sector. There are two methodologies, the CRISP-DM and OSEMN, that is used to explain the data science project life cycle on a high level. The six-phase method framework known as the Cross Industry Standard Process for Data Mining (CRISP-DM) accurately depicts the data science life cycle. On the other hand, the overall workflow performed by data scientists is categorized under the OSEMN methodology.

In our study, we examine both CRISP-DM framework and OSEMN framework and we perform a comparative analysis. We have conducted an empirical study where the experiment was organized into three study cases, each provided insightful results whether which methodology has better model fit and which has a more accurate prediction rate. The study cases suggested that CRISP-DM offers a better performance and accurate approach. All things considered, this research advances our knowledge of best methods, providing practitioners and researchers with direction on which strategy is best suited for their data analysis assignments.

Keywords: *CRISP-DM, OSEMN, framework, data mining, deep learning, machine learning, data science, natural language processing, computer vision*

ABSTRAKT

ANALIZA KRAHASIMORE E METODOLOGJIVE: CRISP-DM PËRBALLË OSEMN DUKE PËRDORUR REGRESIONIN LINEAR DHE ANALIZËN STATISTIKORE

Shameti, Ketjona
Master Shkencor, Departamenti i Inxhinierisë Kompjuterike
Udhëheqësi: Prof. Dr. Betim Çiço

AI ka dhënë një kontribut të rëndësishëm në shumë industri, duke gjetur zgjidhje të reja dhe inovative sfidave të vështira. Megjithatë, aplikimi efiçent i projekteve te AI kërkon një teknikë të strukturuar dhe të kombinuar me qëllimin e vetëm për të qenë të përditësuar me zbulimet më të reja ne këtë sector. Janë dy metodologji, CRISP-DM dhe OSEMN, të cilat përdoren për të shpjeguar ciklin e projekteve të shkencës së të dhënave në një nivel të lartë. Metodologjia e cila zhvillohet në gjashtë faza a njohur si CRISP-DM përshkruan saktë ciklin e shkencës së të dhënave. Nga ana tjetër, rrjedha e punës e performuar nga shkencëtarët është kategorizuar nën metodologjinë OSEMN.

Në studimin tonë, ne kemi analizar të dyja metodologjitë CRISP-DM dhe OSEMN, dhe kemi performuar një analizë krahasuese. Ne kemi kryer një studim empirik ku eksperimenti është organizuar në tre raste studimi, ku secila dha rezultate të vlefshme se cila nga metodologjitë kishte një model më të mirë të parashikimit të të dhënave, dhe cila i parashikoi më saktë të dhënat. Rastet e studimeve sugjeruan se metodologjia CRISP-DM ofroi një performancë më të mirë dhe ishte më e saktë. Duke konsideruar të gjitha gjërat, ky studim na ndihmon të kuptojmë se cilat janë metodat më të mira për të zgjedhur cila prej strategjive ju përshtatet më mirë për projektet dhe studimet e tyre ne lidhje me analizën e të dhënave.

Fjalët kyçe: *CRISP-DM, OSEMN, metodologji, kërkimi it ë dhënave, deep learning, machine learning, shkenca e të dhënave, computer vision, raste studimi*

ACKNOWLEDGEMENTS

There are many people who helped to make my years at the graduate school most valuable. First, I thank Betim Çiço, my major professor and dissertation supervisor. Having the opportunity to work with him over the years was intellectually rewarding and fulfilling. I also thank my cousin who contributed much to the development of this research starting from the early stages of my dissertation work.

Many thanks to my friend, who patiently answered my questions and problems on word processing. I would also like to thank to my graduate student colleagues who helped me all through the years full of class work and exams.

The last words of thanks go to my family. I thank them for their patience and encouragement.

TABLE OF CONTENTS

ABSTRACT iii

ABSTRAKT iv

ACKNOWLEDGEMENTS v

TABLE OF CONTENTS vi

LIST OF TABLES viii

LIST OF FIGURES ix

CHAPTER 1 1

INTRODUCTION 1

CHAPTER 2 4

BACKGROUND 4

CHAPTER 3 7

LITERATURE REVIEW 7

3.1 State of Art 24

CHAPTER 4 27

HYPOTHESIS AND RESEARCH QUESTIONS 27

CHAPTER 5 28

METHODOLOY AND MATERIALS 28

5.1 Datasets 28

5.1.1 The Iris Dataset 28

5.1.2 Life Expectancy Dataset 29

5.1.3 Country Regions Dataset 30

5.2 Libraries 31

5.2.1 Scikit-learn Library 31

5.2.2 Pandas Library 33
5.2.3 NumPy Library 33
5.2.4 Matplotlib Library 34
5.2.5 SciPy Library 35
5.2.6 Seaborn Library 35
5.3 Exploratory Data Analysis (EDA) 36
5.3.1 Univariate Non-graphical: 36
5.3.2 The multivariate non-graphical: 38
5.3.3 Univariate graphical: 38
5.3.4 Multivariate graphical data: 42
5.4 Linear Regression 42
CHAPTER 6 46
EXPERIMENT AND RESULTS 46
6.1 First Study Case 46
6.2 Second Study Case 51
6.3 Third Study Case 58
CHAPTER 7 67
DISCUSSIONS 67
CHAPTER 8 69
CONCLUSION 69
REFERENCES 71

LIST OF TABLES

Table 1. The statistical information for Iris dataset 30

Table 2. The statistical information for Country Region dataset 31

Table 3. The statistical information for Iris dataset 37

LIST OF FIGURES

Figure 2. 1. Four Level Disintegration of the CRISP-DM technique for Data Mining [1].....6

Figure 3. 1. Stages of the CRISP-DM Process Model for Data Mining [1]……8

Figure 3. 2. Summary of the CRISP-DM tasks and results [1].....10

Figure 3. 3. The most pertinent data science and data mining models and techniques have evolved. KDD and CRISP-DM are the grey-colored, "canonical" approaches. [2] 15

Figure 3. 4. The DST map, which shows the search activities' outside circle, innermost ring of CRISP-DM operations. [2] 16

Figure 3. 5. An exemplary path through a data science undertaking. [2] 18

Figure 3. 6. OSEMN workflow [4].....20

Figure 3. 7. Model selection during prototyping phase [7].....22

Figure 3. 8. Most utilized Analytic framework for data scientist [8] 23

Figure 5.3.3. 1 The distribution of sepal, petal length and width……..39

Figure 5.3.3. 2. Distribution among species of Iris Dataset 39

Figure 5.3.3. 3. The representation of numerical values in Life Expectancy Dataset 40

Figure 5.3.3. 4. Distribution of Life Expectancy dataset 41

Figure 5.3.3. 5. Species specific variation in sepal length 41

Figure 6.1. 1. Correlation Matrix for Life Expectancy dataset in OSEMN methodology……...47

Figure 6.1. 2. Correlation Matrix for Life Expectancy dataset in CRISP-DM methodology.....47

Figure 6.2. 1. Predicted vs Actual Sepal Width Values presented in a graph at the modeling part of CRISP-DM methodology…….54

Figure 6.2. 2. Predicted vs Actual Sepal Width Values presented in a graph at the modeling part of OSEMN methodology.....57

Figure 6.3. 1. Predicted vs Actual Life Expectancy Values presented in a graph at the modeling part of CRISP-DM methodology…….62

Figure 6.3. 2. Predicted vs Actual Life Expectancy Values presented in a graph at the modeling part of OSEMN methodology.....65

CHAPTER 1

INTRODUCTION

Artificial Intelligence (AI) has transformed a number of industries, providing creative answers to challenging issues [1]. Nevertheless, an integrated and methodical strategy is necessary for the effective execution of AI initiatives. There are particular difficulties in the creation, administration, and layout of AI platforms. In order to develop solutions that address complicated issues in an unbiased and moral manner, engineers frequently require assistance. Making ensuring the AI framework is accessible and comprehensible is crucial [2]. This, considering the opaque nature of many AI systems, presents a substantial hurdle.

The job of organizing varied groups with expertise in software engineering, data science, machine learning, and corporate expertise falls to AI project leaders. Organizations also have to handle the challenges of incorporating AI technologies into current procedures and operations. This aspect emphasizes how crucial it is to control what stakeholders want and incorporating AI technologies into the current ecology. In addition, managing the demands of stakeholders is essential, since there tends to be greater clarification around what is anticipated of AI as well as what it can actually accomplish [3].

Engineers have to cope with things like accessibility and the caliber of information, choosing models and adjusting, implementation, operation, and system adaptability. Also, they must constantly refresh their abilities and stay current with the quickly changing field of artificial intelligence. This statement highlights how crucial data accessibility and accuracy are to the development of AI systems [4]. These quotations emphasize how difficult it is to create, administer, and construct AI systems. These emphasize the necessity of openness, morality, and skillful customer demand administration, as well as the significance of high-quality information for the advancement of AI.

AI is an area of study that incorporates ideas from mathematics, statistics, computer science, and domain-dependent expertise [7]. Among the fundamental ideas is machine learning, which is the discipline that enables machines to gain insight from information, deep learning, a branch of machine learning that utilizes neural networks with several layers, natural language processing, a technique that enables machines to comprehend what people say, and computer vision, a field of study that seeks to give a computer or other piece of computer code humanoid or superior capabilities [8].

The group follows this technique as they work throughout each stage of the undertaking, from identifying business requirements to putting the AI system into action. Additionally, it highlights the significance of ongoing training and iteration since AI systems must adapt to shifting business requirements and technology advancements [9].

Considering how businesses have been using AI over time and how various methods have been developed using this technology, a significant amount of capital was raised for this kind of project. Evaluating the present state of AI project development requires examining current AI project techniques. To comprehend the development and advancement of AI, it is imperative to conduct a comparative examination of old and new AI design techniques. Conventional approaches, like CRISP-DM, have been in use for many years, show proficiency in initiatives involving data mining and machine learning [10]. But with AI technology developing so quickly, new strategies, as OSEMN, are developing that are more suited to the demands of the present.

Case studies offer insightful information on the practical uses of AI design techniques, enabling examination of the use of these approaches in actual circumstances. Such studies provide a chance to comprehend the difficulties and achievements related to various implementation, emphasizing the versatility of these approaches in many industries [11]. Case studies are becoming an increasingly useful resource for comprehending how AI design techniques are used in real-world scenarios, by offering information on how these methods might be modified and used in various situations to address challenging issues [12].

While well-established and efficacious across multiple sectors, CRISP-DM, on the other hand we have OSEMN, a more modern technique provides a versatile and adaptive strategy, particularly beneficial for tasks involving substantial amounts of information and demand sophisticated machine learning methods [5]. So, in this thesis we are going to compare both the methodologies by implementing a specific project for each, with the same datasets, same logic behind, in order to be able to distinguish which framework is more effective, more efficient and more accurate.

The necessity of ongoing innovation in AI project techniques is highlighted by the contrast of conventional and new procedures, emphasizing the necessity of flexibility and adaptability in selecting the best approach for every engagement.

CHAPTER 2

BACKGROUND

CRISP-DM (Cross Industry Standard Process for Data Mining) project is offering a thorough process model for completing data mining projects. It also needs a set of various abilities and understanding [1]. This is interpreted as so that the good outcome or bad outcome of a data mining project is very much depends on a certain individual or group working on it and its prosperous practice could not always be discussed again throughout the corporate. Data mining is in requirement of a normative strategy that will be in assistance of translating issues related to business into data mining assignments, propose effective data alterations and data mining methods, and offer means or assessing the efficiency of outcomes and recording the work [13]. The CRISP-DM project discussed pieces of the issues by specifying a procedure model that gives a structure for working on data mining projects that is not dependent on the industry division and technology utilized. This procedure model has the purpose of creating big data mining procedures, that have a low cost, are more trustworthy, more recurring, more attainable, and quicker.

The data mining sector is at the moment at a crack along early marketplace and principal stream market. Its financial achievement is however not assured. If the case of early adopters does not have the expected outcomes, they will not fault their incapacity in utilizing data mining appropriately but declare that data mining does not function [14].

Within the marketplace, there exist some assumptions that the anticipation is a push-button technology. Nevertheless, this is untrue, as the majority of the people who practice data mining are conscious. Data Mining is a complicated procedure that need different instruments and various individuals [15]. A positive outcome of a data mining project is based on the correct mixture of good tools and competent analysts. Moreover, the requirement of a sound technique and efficient project oversight is

necessary. This procedure model is beneficial in having a better comprehension and control the exchanges along this complicated procedure [16].

For the marketplace, there will be a lot of advantages in case of a typical procedure model is acknowledged. This model can function as a standard citation point to talk about data mining and will increment the comprehension of important data mining problems by all those who are working with it, in particular at clients' point of view. Nevertheless, the greatest significance remains on making the perception that data mining is a recognized engineering procedure. Customers will experience greater comfort in the case that they are informed a comparable story by various tool or suppliers of services. Practically, customers can attain many sensible anticipations in finding out in what way will the project move forward and what are the anticipations in conclusion [17]. Working with instruments and suppliers of services, it will create a great convenience for them on the task of contrasting various offers to find the best outcome possible. A sensible procedure model will certainly be of great assistance for the distribution of understanding and experience inside the company.

The suppliers will gain from this degree of comfort that the customers experience. The requirement of instructing clients about overall problems of data mining will not be as crucial as it has been. The main emphasis changes according to two options, if data mining should have a usage in general or if it could find appliance to resolve the business-related queries.

Merchants do all they can in order to make their goods more valuable, one example is providing directions throughout the entire procedure or better reuse of outcome and experiences. Those who make possible the service can serve as trainers of employees to a steady degree of proficiency.

The advantages can be known in a lot of forms by those whom execute data mining projects. For beginners, the procedure model offer instructions, assist to build the project, and advising them every stage of the procedure. Also, the more experienced ones can still find use in the instructions for every step of the procedure by ensuring themselves that everything is in order and nothing of great importance has been overlooked. But the function that analysts find of great importance is its use in

documenting and conversing the outcomes. It aids to connect the various instruments and various individuals that have varied set of skills, so as to pertain an efficient and effective work [18].

The CRISP-DM Methodology

This technique can be explained regarding an echelon procedure model, consisting of 4 levels: phases, generic tasks, specialized tasks and procedure instances.

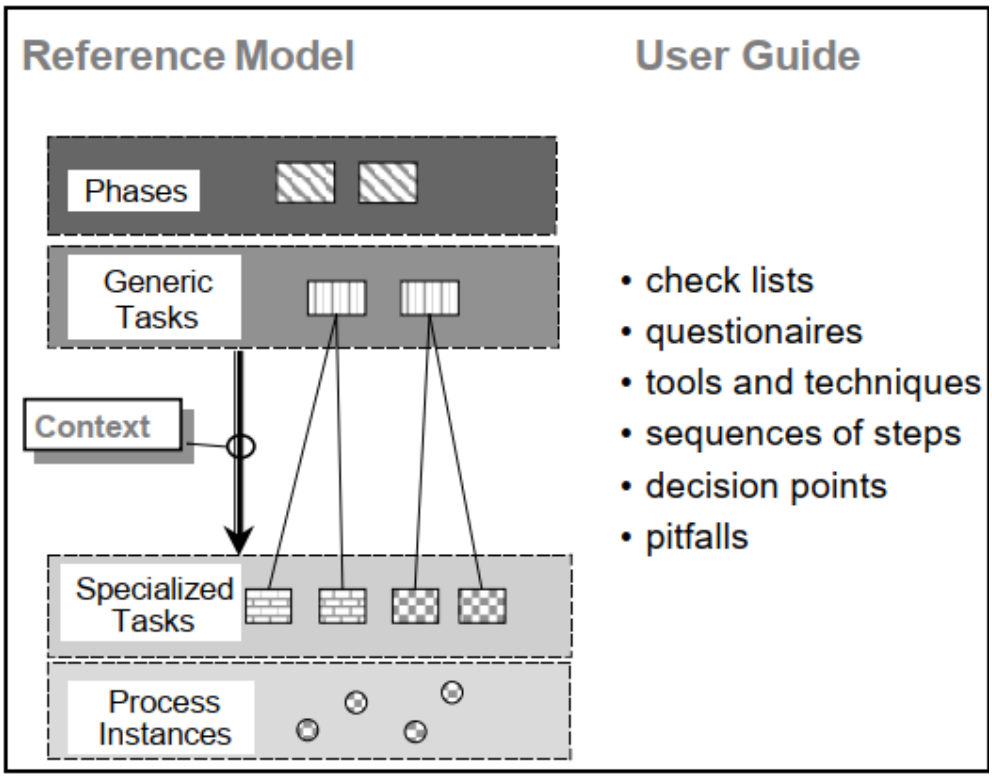

Figure 2. 1. Four Level Disintegration of the CRISP-DM technique for Data Mining [1]

At the highest degree, the data mining procedure is structured into few phases. Each part contains second-level generic tasks. The second level is being called generic, since it is meant to be overall at the right amount as to include every feasible data mining circumstance. This task is created as to perform in the most whole and steady way. Whole is meant to include all procedure of data mining together with every feasible data mining use [19]. While steady is meant to describe the need for the model to be true for however, unanticipated events such as novel modeling methods.

CHAPTER 3

LITERATURE REVIEW

The third tier, the specialized task tier, is where we explain the various ways that activities in the generic tasks have impact, and what activities should take place in certain implications. For instance, one of the model's part of second tier is a general assignment named build model. At the third tier, there is a task named build response model that includes actions that are particular to the issue and to the data mining selected tool [1].

The synopsis of stages and assignments treated as distinct phases executed in a certain sequence shows a perfect flow of things that happen. In actuality, the majority of the assignments are able to carry out in a diverse arrange and sometimes the need arise to revert to prior assignments and do once more specific activities. The CRISP-DM procedure model doesn't try to catch every potential path throughout the data mining procedure since doing so would mean to include an excessively intricate procedure model and even so the predicted advantages are known to be not as profitable as anticipated [2].

The fourth level, ore differently named the process instance level, is a log of activities, choices, and outcomes of a real-world data mining project. An example of a procedure is structured based on the assignments established at the upper tiers, nonetheless, it stands for what took place in reality in a specific participation, as opposed to what occurs generally [20].

The CRISP-DM technique sets apart along the Reference Model and the User Guide. In contrast, the Reference Model shows a fast summary of the stages, assignments and the outcomes, and explains how to act in a data mining project, the User Guide provides us with more thorough advice and suggestions for every stage and every assignment inside a stage and illustrates what goes into a data mining project.

The generic CRISP-DM Reference Model

The CRISP-DM reference design for data mining gives a summary of the life cycle of a project. It provides us with the stages of an engagement, their individual responsibilities, the outcomes for each.

The life span of a data mining project is disassembled in six stages that can be viewed in Fig. 2. The stages' order of occurrence is tolerable. The arrows show just the most significant and regular reliance alongside stages, however, in a certain project, it is restricted to the result that every stage provide, or in another case to find the order in how each stage is going to be executed following [21].

The exterior circle in the picture 2 represents the recurring pattern of data mining in general. Data mining cannot come to an end until a fix is implemented. The knowledge gained while working with the procedure and as well from the implemented resolution may result in fresh, sometimes more targeted business inquiries. Later data mining procedures will use the prior encounters and of course will be very beneficial.

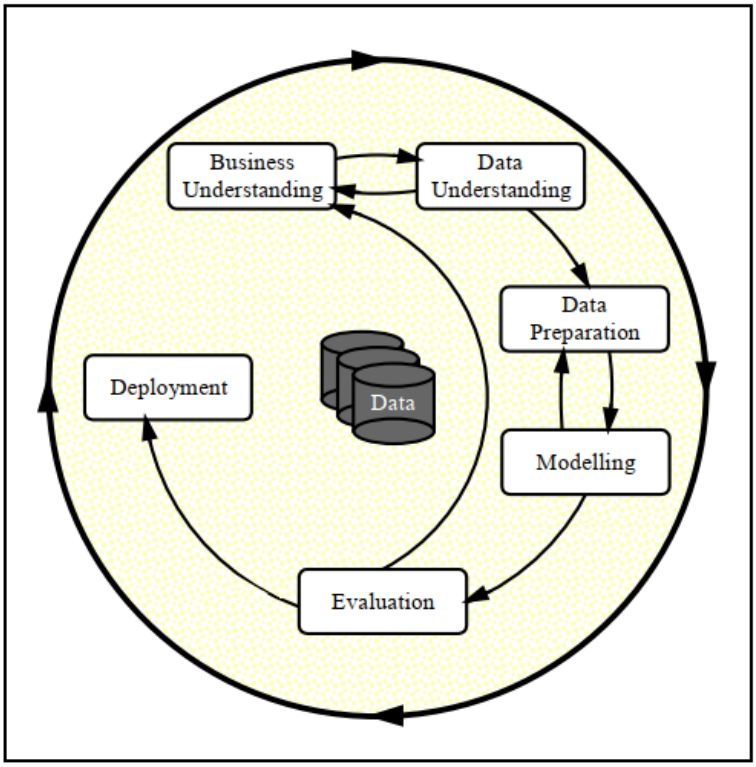

Figure 3. 1. Stages of the CRISP-DM Process Model for Data Mining [1]

In the subsequent, we provide a quick overview of every stage as shown in Fig. 3:

- **Business Understanding**

This first stage emphasizes comprehension of the engagement's goals and specifications regarding the standpoint of a business, and after that transforming this understanding to a data mining issue definition, as well as an initial engagement scheme made to accomplish the goals.

- **Data Understanding**

The data comprehension stage begins with a first gathering of information and continues with actions to familiarize oneself with the information, to be able to determine what are the issues related to the caliber of the information used, to find out initial perceptions of the information, or to identify intriguing subgroups in order to create theories about the concealed details [1].

We can distinguish a strong connection alongside Business Understanding and Data Understanding. The way how the data mining issue is proposed and the engagement plan must have a minimum of a certain level of comprehension of the information at hand.

- **Data Preparation**

The data preparation stage encompasses all building procedures of the final dataset (information which is going to be put into the modeling instruments) taken from the original unprocessed information. Data preparation tasks will be executed several times in the majority of cases, and also is worth mentioning that they will not be executed in the designated sequence. Assignments consist of table, documents, and choice of attributes, data cleaning, creation of novel characteristics, and information conversion for modeling software [1].

- **Modeling**

In this stage, different modeling methods are chosen and utilized, and their settings have been adjusted to ideal levels. Usually, in order to solve a certain kind of data

mining issue, there could apply a number of methods. Some of them need certain data formats.

We can distinguish a strong connection alongside Data Preparation and Modeling. Usually, one becomes aware of data issues when dealing with arranging or even if someone has inspiration for creating new information [22].

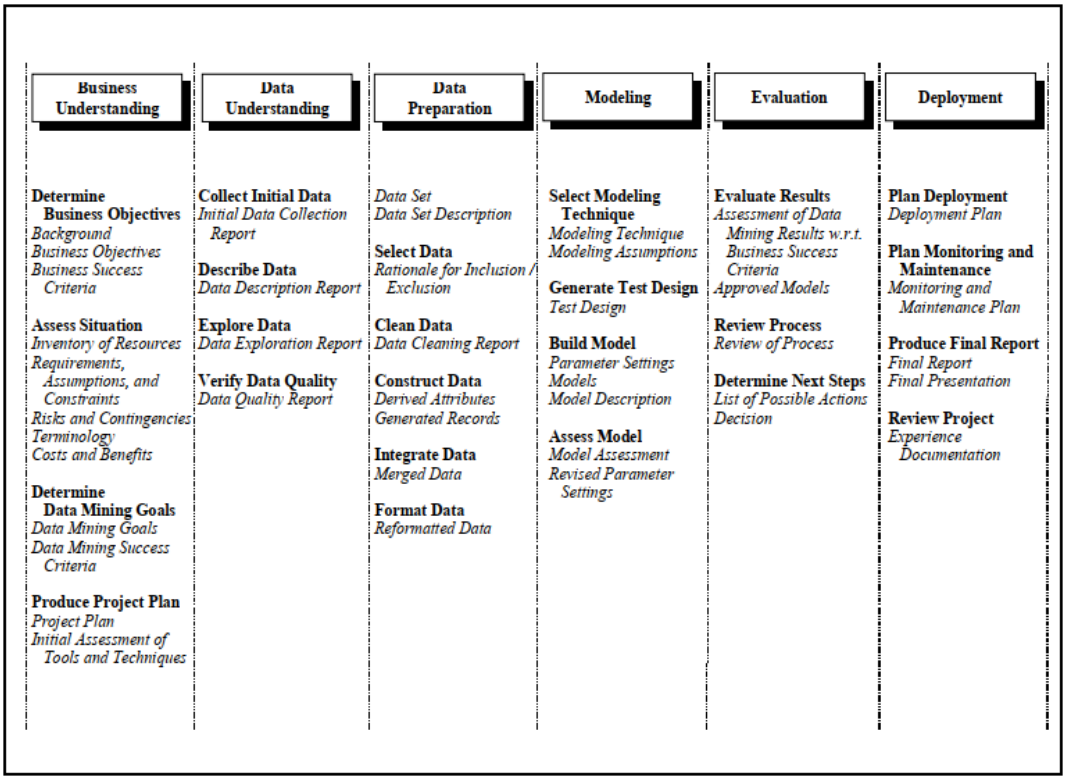

Figure 3. 2. Summary of the CRISP-DM tasks and results [1]

- **Evaluation**

At this part of the methodology, there seem to be constructed some models with superior excellence, from the standpoint of data interpretation. It is crucial that we fully assess the model and analyze the actions taken in order for the model to be built, then we can move forward to final deployment [23]. Then we can sure that it successfully accomplishes the corporate goals. One of the main goals is to ascertain whether there are several significant business problems that has not been given enough thought. In the final step, a choice about the application of data mining outcomes ought to be made.

- **Deployment**

Model development is typically not the final phase of the undertaking. Typically, the acquired understanding will require planning and displayed in a manner that the client is able to utilize it. Based on the specifications, the deployment step is very easy and can be compared to producing a report, but could also be as difficult as putting in place a replicable method for data mining. Frequently, is the user that will execute the deployment phase, not the data analyst. Anyway, it has a great significance to comprehend initially what steps are necessary to be executed to be able to genuinely utilize the produced models [24].

Reaction simulation is a method to increase the efficacy and efficiency of marketing mail campaigns. It makes it possible to raise the pace of response while lowering a campaign's expenses. By extending the understanding we currently possess regarding our potential using data mining techniques, we can estimate the probability of prospective clients to respond to our correspondence.

Regarding a mailing operation, primary goal of reaction modeling is to limit the correspondence to recipients who are members of the designated collective. Other than this, reaction modeling aids in finding out and to comprehend our opportunities. Thus, our constant goal is to obtain an intelligible, practical profile of the intended audience as an additional prerequisite [25].

To enable the training of models based on our objectives we make use of historical data that include data on the intended course of action, for example, possess a quality that determines if a client changed his mind about the car that he initially thought of buying. We get this information from client database, from generic questionnaires, or purchase from address finders.

The objective of the engagement in this case was the formation of a systematic procedure that marketers may consistently carry out with less experience in data mining and insufficient time to try out various strategies. The first round of case studies is completed by the main project group, creates and keeps up the procedure, develops

the marketing personnel, who will thereafter be supported by them with more difficult application situations that deviate from the norm [26]. The group working on the engagement comprises both knowledgeable data miners and experts in marketing.

The first round of case studies centered around purchase campaigns, for example choosing potential customers that are probably going to get a Mercedes for the first time, from an address list is difficult. Even if this seems like usage of data mining in books, numerous negative ramifications exist.

To begin with, the procedure is not as steady and consistent as one could anticipate. Numerous elements exist, such as the information at hand or the state of the marketplace, these in some ways distinguish one reaction modeling initiative from the others. Things become worse if we wish the procedure to be used in many European nations with linguistic variations, customs, legislation, and a scenario of competition [27].

Thus, the following conundrum arises: In the first place, it is plainly impractical and, ultimately, undesirable to provide a thorough and in-depth explanation of the procedure. This is because of how complicated due to the numerous unknowable elements and unforeseen circumstances that will never go away. We encountered the fact that inside the same nation the situation may be essentially different despite the campaign. Conversely, though the reverse extreme, a very elevated level summary similar to the standard CRISP-DM, is also not a remedy. Despite covering the entire procedure and it is beneficial for those with experience, it is inappropriate given the type of users when migrating into routine business operations, one is faced with. What emerged was a description of the method that ought to direct the consumer as much as feasible, however, concurrently, allow him to manage challenging unforeseen circumstances.

In this part we provide a brief explanation among the most frequently utilized and mentioned data mining and the acquisition of understanding techniques, giving a summary of each's development, foundation and essential traits. For an in-depth explanation of these techniques we use the complete procedure of data-driven learning of knowledge, encompassing the methods used for storing and retrieving information,

methods for sizing techniques to enormous datasets and continue to function effectively, how findings can be visualized and understood, and how the general exchange between humans and machines can be designed and provided with assistance, and data mining as one action in the procedure, converting appropriately preprocessed data into designs that can then get transformed into useful and practicable understanding. Still, data mining is frequently employed as an alternative term for KDD, and we won't make any distinctions amid the two interpretations in this work [28].

CRISP-DM can be observed as the standard methodology where the majority of the ensuing suggestions have changed (regarding the data mining and data science procedure creation). It expands and explains the actions inside the initial KDD suggestion into six actions: Business understanding, Data understanding, Data preparation, Modelling, Evaluation, and Deployment. Numerous procedure prototypes and techniques were created approximately at the year 2000 using CRISP-DM as a foundation, but with different goals.

There are a few more pertinent methods as well, not directly connected to the KDD assignment. The 5 A ’s Procedure, that was initially created by SPSS, incorporated a "Automate" phase previously that aids novice consumers to fully mechanize the procedure of DM submitting an application previously established techniques for fresh information, yet, it lacks stages to comprehend the goals of the business and to assess the data caliber. An alternative strategy that attempts to help the users in the DM procedure. Each of these had some impact for CRISP-DM. The KDD Roadmap is an additional strategy, an incremental data mining technique that as the primary input presents the “resourcing” task, comprising the incorporation of databases derived from several references to create the functional database [29].

It provides a graphic representation of how various data mining procedure models and approaches have evolved. As seen in the illustration, the darts show that CRISP-DM includes concepts and ideals from the majority of the previously listed techniques, while also developing the origin of numerous subsequent ideas. CRISP-DM is still regarded as the most comprehensive approach to data mining regarding the gathering of the requirements of commercial initiatives, and is currently the most utilized

procedure for DM initiatives based on the KD nuggets surveys kept in 2002, 2004, 2007 and 2014. To put it briefly, CRISP-DM is regarded the informal benchmark for statistics, data mining and initiatives in data science.

But data science is currently a lot more frequently employed phrase than data mining with relation to the pursuit of understanding. There appear to be two major perceptions where the phrase is employed: (a) the science OF data; and (b) utilizing scientific techniques with TO data. From the first vantage point, data science is thought of as a topic from academia that investigates information in all of its forms, along with techniques and formulas to modify, examine, present, and improve data. In terms of methodology, it is similar to statistics and computer science, integrating conceptual, algorithm-based and practical research [30]. From a different angle, data science encompasses both the academic and industrial domains, obtaining worth derived from information utilizing scientific techniques, like statistical hypothesis testing or machine learning. In this case, the focus is on finding solutions for specific to a domain issues using data as a guide.

Data are utilized to create models, design elements, and usually deepen comprehension of the topic. Should we wish to differentiate between these two perceptions, therefore the first theoretical information might be referred to as scientific; and data science in practice, which is the second. The latter is essentially what we are focusing on in this study and from now on, we will refer to it as "data science" in this practical perception. The primary distinction that we note amongst data mining 20 years ago as well as data science nowadays has objectives and focuses on the procedure, whereas the latter is focused on data as well as investigating as shown in Fig. 4. Originating from the purpose-driven viewpoint, CRISP-DM is centered on procedures and various assignments and functions inside those procedures [31].

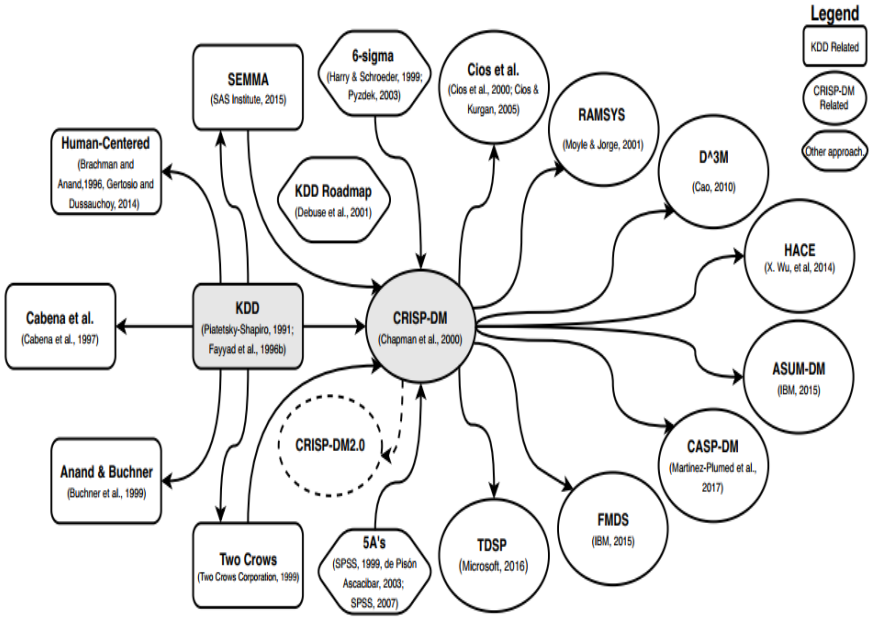

Figure 3. 3. The most pertinent data science and data mining models and techniques have evolved. KDD and CRISP-DM are the grey-colored, "canonical" approaches. [2]

It sees information as a component needed to reach the objective. Put another way, the procedure is central to the data mining viewpoint. On the other hand, information is central to modern data science: we have faith or suspicion that this information has worth; how can we access it? What potential steps could we take with the information in order to make use of and uncover its worth? Stepping aside from the procedure, the approach grows increasingly investigative and less normative: actions you may perform with information as opposed to actions you ought to take with data [32].

Using the analogy of "mining" once more: if mining information is similar to extracting valuable metals, Data science is similar to mining: looking for valuable metal resources where viable mines might be found. A scouting procedure like this is essentially investigative and may involve a few of the subsequent pursuits:

Investigation of objectives: identifying organizational goals that can be fulfilled via data-driven approaches;

Investigating information sources: finding fresh and useful resources for information;

Data value investigation: determining the possible worth that may be derived from the information

Investigating results: connecting data analysis findings to corporate objectives

Study of storylines: obtaining worthwhile narratives from the information, for example either written or visual

Product investigation: figuring out how to create a solution or application using the benefit that can be taken from the information and provides users and clients with something fresh and beneficial

The field of study determines the sequence of tasks in data science in addition to the choices and findings made by the data scientist. For instance, after obtaining inadequate outcomes from information value discovery depending the information provided, additional information source investigation may be required. Conversely, if no information is provided hence investigating sources of information will occur ahead of investigating data values. These two actions can occasionally be repeated and at times none of them is necessary [33].

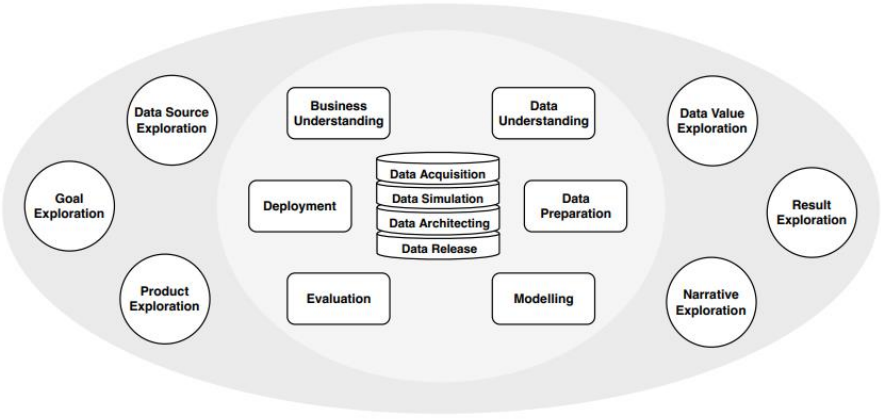

Figure 3. 4. The DST map, which shows the search activities' outside circle, innermost ring of CRISP-DM operations. [2]

Engagements using data science are undoubtedly about more than just discovery, and also have additional segments that are driven by objectives. The CRISP-DM strategy six basic phases from business comprehension to deployment remain applicable as

they are today as shown in Fig. 5. Yet, it is typical for data analysis initiatives to simply show fragmentary trails using CRISP-DM. For instance, occasionally actions outside preparing the data are not necessary, since the information prepared represent the engagement's end result.

Information that has been taken from several sources, blended and cleaned material may be purchased or released for a number of uses, or put into a data center to be queried using OLAP. The stages of CRISP-DM are frequently halted by additional research endeavors, anytime the data scientist wishes to look for further details and fresh concepts. Therefore, we believe that an excellent data science endeavor will go across a space along a trajectory similar to the one shown in Figure 5. Unlike the CRISP-DM model, this one lacks arrows, since there is no set order in which the tasks must be completed. The manager or leadership of the project are in charge of determining the next course of action, depending on the data at hand, considering the outcomes of earlier initiatives [34]. Despite the fact that the area includes every CRISP-DM stage, these don't always happen in the prescribed sequence, because investigative actions are placed between goal-driven operations, and these occasionally establish new objectives or offer fresh information.

Counterfactual prediction has to take subject expertise into account not just to define the objectives or queries that need to be addressed, but also to locate or produce the information sources, but also to properly outline the framework's fundamental structure. In CRISP-DM, this assignment along with others involving causal inference proceed smoothly however, specific expertise becomes essential. The organization's comprehension stage, on the other hand, strengthens its starting point in these conditions, since this is exactly where the domain expertise should be applied and must be transformed into the visualizations and queries required for the next processes [35].

Yet, data science needs to use the data more actively within the causal inference paradigm. Data is more than just a system's input. This points to a more iterative method for which producing new information may be necessary, for example, by running simulations on the produced or obtained information or by conducting trials that are randomized, employing visual representations to represent the specialist's causal expertise in addition to additional domain expertise or retrieved patterns. It is

challenging to incorporate each of these activities into the CRISP-DM model and might necessitate developing fresh creative tasks for modeling and information collection.

Another pertinent aspect in which CRISP-DM appears to be lacking is when considering “data-driven products”, like a smartphone application that gathers data from people's whereabouts and gives other people travel recommendations, based on their tendencies. The information and expertise gleaned through the information make up the product. Twenty years ago, this viewpoint was uncommon, but it is now widely held. Additionally, the information may now be used for a variety of purposes, even outside of the setting or area in which they were originally obtained. These days, certain systems' enormous and complicated data sets imply that processing the information necessitates significant technical effort related to network and curation [36].

Put otherwise, the CRISP-DM approach treated the "information" like an immovable disk cylinder midway through the procedure, but we wish to draw attention to the activity surrounding this disk, despite the incorporation and compilation of information. Considering the range of options for utilizing the information that you or others have provided, we take into consideration the subsequent requirements for data management for your advantage and for the use of others.

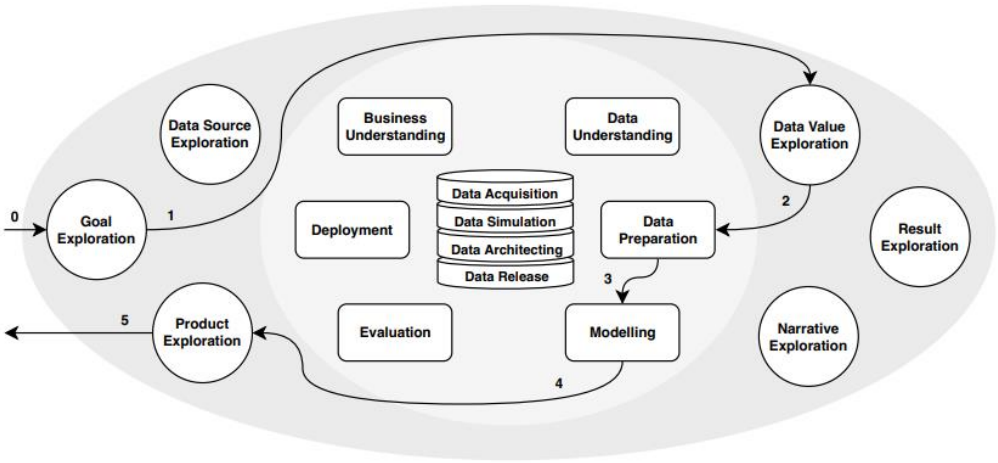

Figure 3. 5. An exemplary path through a data science undertaking. [2]

The OSEMN method is a set procedure and extensively used model of arrangement of investigation in the Data Science domain. It resolves the issues with Data Science/Analytics on a big level as shown in Fig. 6. The procedure for obtaining and changing information that must be put in order, adequately prepped and preprocessed. Utilizing the OSEMN procedure offers a well-defined sequence of tasks: Obtain, Scrub, Explore, Model the data and interpret the data.

By doing these actions, it is possible to think ahead and coordinate the entire workflow, beginning with the collection of data to the outcomes display of the information analysis in an environment for software that has been specifically created. The workflow is well-structured and well-organized. It is composed of multiple logical subsequent actions by means of which the initial objectives are fulfilled [37].

The term analyzers that are machine learning oriented are a subset of models for supervised machine learning, in which some completely annotated trained data are required for the classifier to be developed prior to its application to the real categorized work. Typically, the training information resides in an alternative section of the distinctive, independently hand-labeled information. Following appropriate training they are applicable to the real test information. Naive Bayes is a type of classifier while SVM is a specific type of vector space classifier. It is a type of classifier that needs that the written word records ought to be converted into feature vectors prior to their application in categorization. Text files are typically converted to vectors with several dimensions or possess many dimensions. The whole categorization conundrum is subsequently categorizing each text file. This kind of big range analyzer is what it is [38].

Natural Language Processing

NLP is an area of technological study, AI, as well as logical worried about the exchanges among human linguistics and machines. This method makes use of the openly accessible library, which offers an appreciation for opposite feelings for every phrase that appears in the written work. Within this vocabulary source every term that appears in the written material has a meaning related to three ratings expressed in numbers object, positive and negative, explaining the corresponding interpretations of

the phrase. The outcomes obtained are combined to determine each of these ratings using eight ternary classifiers.

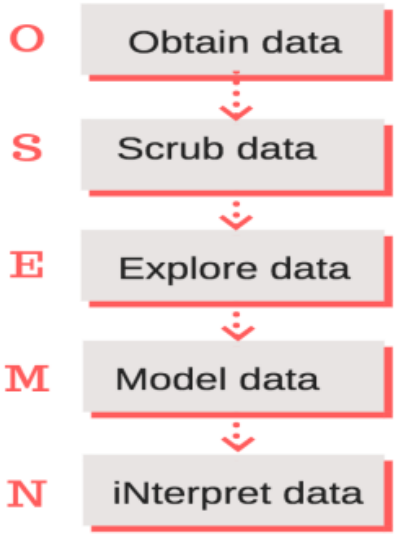

Figure 3. 6. OSEMN workflow [4]

Obtain data

Sensors provide the information. This information is utilized to keep an eye on the conditions and circumstances. External sensors are positioned in various places that creates a precise notion. Devices are arranged based on particular information required by the individual using them. A collection of interconnected detectors for a standard microprocessor, therefore creating a node. A single node gathers particular kinds of information permitting many nodes to operate concurrently, with an almost infinite quantity of nodes. Based on the amount of the apiary and the network's demand, there exist multiple varieties of microcontrollers.

The initial thing they do is gather and acquires the information. Collecting data is to compile information from numerous places, this initial action is vital because If there is no data, then what is the way to proceed with the undertaking. We use database queries to retrieve information for extraction. We get information in a Microsoft Excel spreadsheet to collect or retrieve the information, which we employed to transform it into a collection of information or structure that is useful, thus we are transitioning

from unsupervised to supervised data. The alternative method they are employing to gather information is utilizing web-based APIs like Twitter [39].

Data Exploration

The procedure of exploring information that need to investigate for it before moving on to MI and AI is called data exploration. Initially, we must examine all of the information features. Numerous types of information exist, such as categorical information, periods, statistical information and so on. We do, however, have categorical data. Vibrates can fall into any group and all we need to do is assign them to the appropriate group. The computation of descriptive statistics is going to be the following action that is collecting characteristics and evaluating significant factors. Evaluating significant factors is frequently carried out using connection relationship [40].

Data Modelling

This phase of the data science project life cycle is the most interesting one. Once more, before you get to this point, remember that the cleaning and investigation stage is essential for this procedure to have any sense. One of the initial steps in data modeling is to reduce the size of the information's dimensions. Not every one of your traits or ideals is necessary to accurately describe what you represent. In this stage, we use several algorithms, and we categorize the information based on those techniques. One of the techniques utilized might be naïve byes. We could select the two groups among the two languages English and French based on the classification and transcribe those groups' chirps in both tongues and verify the precision percentage for each category and contrast the two tongues' rates of precision.

Apart from the prediction of the outcome and expulsion the objective of this phase can involve grouping together information to comprehend the reasoning underlying those groups as shown in Fig. 7. For instance, you are interested to organize the users of your online store to comprehend how they behaved on your site. Therefore, to do this, you must be able to recognize sets of information values using techniques based on the

clustering algorithm such as k-designates or create a premonition using regression analysis such as linear or logistic regression as shown in Fig. 8 [4].

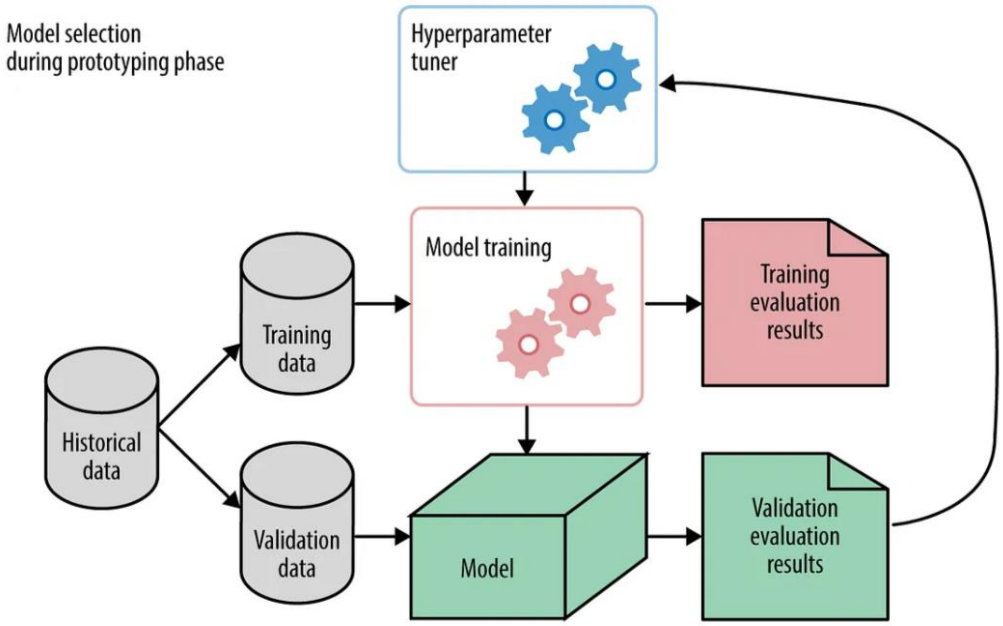

Figure 3. 7. Model selection during prototyping phase [7]

Data Interpreting

We have reached the last and most significant stage, which is data interpretation. This just relates to how the data is presented, presenting the findings in a manner that can respond to the commercial inquiries you made at the outset of the undertaking, in addition to the useful insight obtained from data science. We will provide or symbolize those tongues' chirps in this stage, whose precision rate exceeds that of the other the tongue's chirp [2].

This implies that we will provide an explanation of the outcome that yields a useful conclusion. One way to express this result is by utilizing pie chart. Presenting the results in a manner that is ancillary to our organization is highly crucial, if else our customers wouldn't have been able to access it. We needed a different expertise in place of technical expertise also, which is the capacity to narrate an understandable and relevant narrative as shown in Fig. 9.

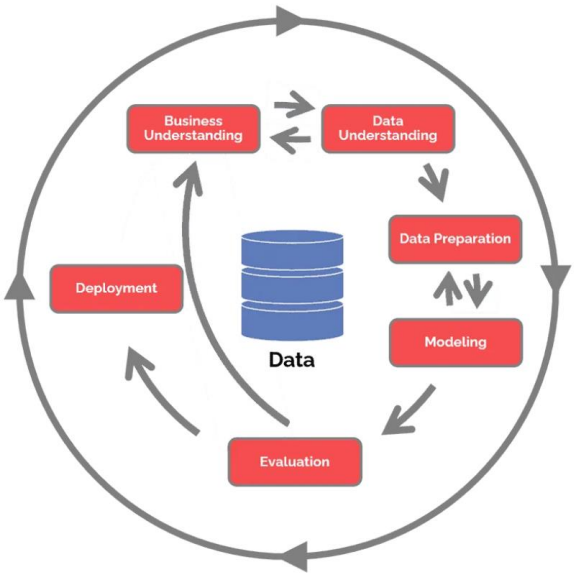

Figure 3. 8. Most utilized Analytic framework for data scientist [8]

Selecting a uniform procedure for processing data is a blend of anticipated efficiency and ease of use. The focus on established procedures is a work attitude that turns attention from the actions to the outcomes since tasks are completed in an organized way to produce value for the final product. The simple to follow and rationally sound procedures of the pipeline for data processing, strengthened with further guidance, observations and model texts, guarantee the actions are carried out and the attainment of the desired outcomes similarly by the various players. An increased sophistication in terms of harmonizing and standardizing procedures at various points in time is made possible. Inadequate awareness-related mistakes are prevented.

Among the many crucial aspects of regulated work procedures is the fact that they clearly define not just the order not to mention the obligations. Every step of the procedure makes it apparent what is required, who is going to be asked to do a specific task and to whoever the result ought to be given. The goal of upcoming research is to achieve visibility at every level of the undertaking that it will make identifying errors simple and, if necessary, quickly return to a particular procedure phase [17,20].

3.1 State of Art

We have analyzed four papers that are similar to the topic that we choose. The title of the first paper is “CRISP-DM: Towards a Standard Process Model for Data Mining” written by Rüdiger Wirth and Jochen Hipp. In a reaction modeling program undertaking, they implemented and evaluated the CRISP-DM approach. The project's ultimate objective was to define a procedure that can be effectively and dependably reproduced by many individuals and modified for various contexts [1]. The initial endeavors were carried out by seasoned data mining professionals; subsequent projects will be handled by individuals with less computational expertise and minimal opportunity to try out various strategies [1]. It turns out that the CRISP-DM technique, with its differentiation between general and specialized process models, offers the flexibility and structure required to meet the requirements of both groups. The general CRISP-DM procedure paradigm is helpful for documentation, planning, and communication both inside and outside the project team. Even seasoned individuals can benefit from the general check-lists. The overall procedure model offers a great starting point for creating a customized process model that outlines the actions to be done in detail and offers helpful guidance for each stage. It turned out that the standard procedure framework was helpful for planning and documentation, as they had anticipated [1]. But it turned out that the framework was considerably more useful for interacting inside and outside of the project than they had first thought. But occasionally, they ran across issues as a result of not adhering to the framework. Sometimes they decided we could get by without the laborious planning and documentation duties, so we bypassed them. However, they ultimately certainly spent a longer period than they would have if they had done the appropriate amount of specific planning [1]. Even though it is emphasized multiple times in the CRISP-DM documentation that the stages and operations are not meant to be exactly ordered, their decision makers were unavoidably left with this impression due to the process model's evident and straightforward presentation.

The title of the second paper is “CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories” written by Fernando Martınez-Plumed, Lidia Contreras-Ochando, Cesar Ferri, Jose Hernandez-Orallo, Meelis Kull, Nicolas

Lachiche, Mar´ıa Jose Ramırez-Quintana and Peter Flach. In this study, they examine whether CRISP-DM is still appropriate for use in data science initiatives and under what conditions [2]. They contend that the procedure model approach is still generally valid in the event that the undertaking is goal-directed and procedure-driven. However, as data science endeavors get more preliminary, a more adaptable architecture is required because there are more possible project paths. They offer some suggestions for the general structure of such a trajectory-based framework and how data science projects (goal-directed, preliminary, or data administration) can be categorized using it [2]. They compare these seven real-world examples—where investigative efforts are crucial—with 51 applications taken from the NIST Big Data Public Working Group. They hope that this classification will aid in the planning of projects concerning time and cost parameters. Since data science is still in its infancy, there are still a lot of unanswered issues about its foundation and methods [2]. They tried to take a more bottom-up strategy to these topics than other authors who have used a top-down strategy, examining how a technique that is widely acknowledged to be beneficial in the area of data mining may be expanded upon to take into consideration the much deeper context of data science. Therefore, they regard this as a part of a bigger, continuing discourse, and they believe that this viewpoint will be seen as a useful addition [2].

The title of the third paper is “OSEMN Approach for Real Time Data Analysis” written by Kajal Kumari, Mahima Bhardwaj and Swati Sharma. The framework that they use to gauge the precision of both languages' chirps is the information analysis system [3]. The fact that this project is new is what matters most. They can argue that sentiment analysis is only a small portion of it, but first they must define sentiment study and categorization. Sentiment classification is therefore a technique to examine the private information contained in the chirps or data and then retrieve the viewpoint [3]. Chirps examine how data is extracted from people's feelings and judgments about various objects. When making decisions, seeking advice from others can have a significant impact on the convenience of users or consumers. This is since they are making decisions about events, products, e-commerce, and other things. The methods for analyzing chirps operate at a certain level, the document stage [3]. The goal of this work is to analyze an approach for sentiment classification at a powdery, specifically

in phrases where the three categorization names positive, negative, and neutral indicate the polar character of the chirps or words. Because real-time data is being used in this approach, they will be able to compare the findings and determine the accuracy of both languages through this procedure. To increase the precision of the results, they are employing thousands of chirps [3].

The title of the fourth paper is "OSEMN Process for Working Over Data Acquired by IOT Devices Mounted in Beehives" written by Kristina Dineva and Tatiana Atanasova. Methods for gathering, organizing, analyzing, analyzing, modeling, and understanding IoT data are a significant problem and a significant obstacle for many researchers [4]. The launch of the OSEMN, a standardized model of work, regulates the problem-solving procedure. A consistent procedure is required for beekeeping, a subsector of the agricultural sector, to use data from sensors housed in beehives. Important information about the behavior of individual bee colonies is obtained after appropriate data processing, which aids in the identification of connections between the various occurrences and the factors that trigger them [4]. This article's goal is to explain the OSEMN model and how beekeeping uses it. Beekeepers can increase and incorporate new revenue streams, save expenses, and improve industrial efficiency by incorporating modern technologies into their operations [4]. Predicting occurrences and establishing a correlation among the information being analyzed and activities taking place in beehives is the ultimate goal of gathering, purging, examining, modeling, and comprehending the IoT data from the beehives. Standardized work procedures are crucial because they clearly outline the steps involved as well as the roles that each person will play [4].

In our work we have directed our focus on a direct comparison of the CRISP-DM and OSEMN data mining methods utilizing real-world application. Our goal is to ascertain whether strategy performs better in terms of efficiency, accuracy, and prediction by carefully analyzing both.

CHAPTER 4

HYPOTHESIS AND RESEARCH QUESTIONS

We have raised some hypothesis and research questions that our study aims to answer at the end of the paper.

H_0 – The CRISP-DM framework has better model fit than OSEMN framework.

- **RQ1:** Has the CRISP-DM methodology performed better when considering the metrics of R-squared and MSE?

H_1 – The CRISP-DM methodology has a more accurate prediction rate than OSEMN methodology.

- **RQ2:** What does the graph representation for both indicate when comparing the prediction accuracy?
- **RQ3:** What does the metrics suggest us when comparing both methodologies regarding the prediction rate?

CHAPTER 5

METHODOLOY AND MATERIALS

5.1 Datasets

5.1.1 The Iris Dataset

The first dataset used is Iris Dataset. In 1936, British statistician and biologist Ronald Fisher published a paper titled "The use of multiple measurements in taxonomic problems," which established the dataset. The four characteristics (sepal and petal length and breadth) of fifty specimens from three different iris species (Iris setosa, Iris virginica, and Iris versicolor) are included in the Iris Dataset. To categorize the species, a linear discriminant model was constructed using these metrics. Yet, because there just two clearly distinct groups in the data set, using it for cluster analysis is not very frequent. Because of this, the data set serves as a useful illustration of the distinctions between supervised and unsupervised methods for data mining.

```
from sklearn.datasets import load_iris
iris=load_iris()
print (iris.DESCR)
iris.feature_names
```

We import the load_iris function from the Scikit-learn library that contains a number of different datasets. We just need the iris one so we import only it, and the function makes possible to transfer the dataset into our Python Environment. Then we assign this function to the variable iris and output the description and the names of the tools. We will provide a screenshot of the output.

However, when projected onto the nonlinear and branching main element, all three species of iris may be distinguished from one another. The nearest graph approximations the data set, penalizing it for having too many nodes and for twisting

and extending. Subsequently, the "metro map" is built. The nearest cluster is projected with the data values. A pie chart of the anticipated values is created for every node. The number of expected spots and the pie's area are proportionate. The graphic on the left side makes it evident that almost all of the samples from the several Iris species are associated with distinct nodes.

5.1.2 Life Expectancy Dataset

While many studies on aspects influencing life expectancy, considering death rates, income layout, and demographic characteristics, have been conducted in past times. It was discovered that the relationship between the index of human advancement and immunizations was not previously considered. Furthermore, a portion of previous studies examined multiple linear regression using a year's worth of information for each nation. This provides incentive to address the two previously mentioned problems by developing a regression model based on multiple linear regression and a combined effect approach, considering data for all nations from 2000 to 2015. Vital vaccinations including those against polio, diphtheria, and hepatitis B will also be considered. To put it briefly, the research will concentrate on elements connected to immunization, mortality, economics, society, and other aspects of health.

```
import pandas as pd
life= pd.read_csv('/content/data/Life_Expectancy_Data.csv',index_col='Country')
life.head()
```

We import the pandas library and use one of its functions read_csv in order to upload the excel file where our dataset is. Inside the parameters of the function we put the path where excel file of life expectancy in located in the directory of the laptop. Then we only output the first few rows of the file just for a simple representation.

Table 1. The statistical information for Iris dataset

Contin-ent	Year	Status	Life_ expectancy	Adult_ Mortality	infant_ deaths	Alcohol
Asia	2015	**Developing**	**65**	**263**	**0.01**	**71.279**
Asia	2015	**Developing**	**65**	**263**	**0.01**	**71.279**
Asia	2015	**Developing**	**65**	**263**	**0.01**	**71.279**
Asia	2015	**Developing**	**65**	**263**	**0.01**	**71.279**
Asia	2015	**Developing**	**65**	**263**	**0.01**	**71.279**

The content of this dataset are the columns: Country, Continent, Year, Status, Life Expectancy, Adult mortality, infant deaths, alcohol, percentage expenditure, hepatitis-B, measles, BMI, under five deaths, polio, HIV/AIDS, total expenditure, diphtheria, GDP, population, thinness, income composition of incomes, schooling. Each row has information for every column, there are 156 countries that contain the information expressed in the table. This dataset contain information from year 2000 till 2015.

5.1.3 Country Regions Dataset

This dataset contains information about 249 countries. Its column are: name of the country, alpha-2, alpha-3, country code, iso-3166-2, region, sub-region, intermediate-region, region code, sub region code and intermediate code. Each country contains specific information regarding the content of the columns provided.

```
df = pd.read_csv('/content/data/country-regions.csv.csv')
df.head()
```

We upload this excel file as well, that contains the explained information. The path of the directory where the csv file is stored, is put inside the parameters part of the function. This way we can access the information provided by this excel file. Then we print the initial rows along with the column names.

Table 2. The statistical information for Country Region dataset

Name	Alpha-2	Alpha-3	Country Code	Iso_3166-2	Region	Region Code	Sub-Region Code
Afghanistan	AF	AFG	**4**	**2:AF**	**Asia**	**142**	**34**
Åland Islands	AX	ALA	**248**	**2:AX**	**Europe**	**150**	**154**
Albania	AL	ALB	**8**	**2:AL**	**Europe**	**150**	**39**
Algeria	DZ	DZA	**12**	**2:DZ**	**Africa**	**2**	**15**
American Samoa	AS	ASM	**16**	**2:AS**	**Oceania**	**9**	**61**

5.2 Libraries

5.2.1 Scikit-learn Library

A Python package for statistical modeling and algorithms for machine learning is called Scikit-Learn, or sklearn. A variety of machine learning models for regression, classification, and clustering can be put into practice, and these models can be analyzed statistically using tools. Additionally, it offers capability for aggregation approaches, obtaining features, choosing features, dimensionality reduction, and built-in datasets. Numerous built-in datasets, like the iris, home price, and diabetes datasets, are included with Scikit-learn.

from sklearn.datasets

Here, we have provided some characteristics of this library.

The ability to divide the dataset for testing and training purposes was made possible by Sklearn. Divide the dataset in half to make it possible for an objective assessment of prediction accuracy. When the outcome factor is periodic and has a linear relationship with the dependent variables, this supervised machine learning model is employed. Similar to linear regression, logistic regression is a supervised regression

algorithm. The outcome of the parameter is categorical, which is the only distinction. One effective tool for both classification and regression issues is the decision tree. It makes judgments and forecasts the result using a model resembling a tree.

Many or even several hundred decision trees are employed in the Random Forest bagging technique to create the representation. Random Forest is applicable to issues involving both regression and classification. It can be applied to forecast illnesses, detect forged documents, and categorize those seeking loans. The acronym for eXtreme Gradient Boosting is XGBoost. Gradient boosted decision trees are implemented with excellent results using this boosting method. The primary characteristics of XG-Boost are its ability to automatically manage information that is lacking, assistance with normalization, and overall significantly higher accuracy than previous versions.

The supervised machine learning process known as the "Supervised Vector Machine" involves plotting every information point in an n-dimensional space, where n is the total amount of distinct characteristics in the dataset. Subsequently, we carry out categorization by identifying the hyperplane that effectively separates the categories. A table that describes how well categorization methods operate is called a confusion matrix. The sorting algorithm's forecasts are examined using a classification report. It offers the different metrics (accuracy, precision, recall, and f1-score) that allow us to assess how effectively our model is doing.

An uncontrolled machine learning method called K-Means clustering is employed to address classification issues. An approach that lacks an indicator or outcome factor in the dataset is said to be unsupervised. During the clustering process, the dataset is divided into multiple categories, or clusters, according to shared traits and attributes. Another unsupervised clustering algorithm that creates groups based on data point similarity is called DBSCAN. A dimensionality-reduction technique called principal component analysis is employed to shrink huge datasets to a size where the shortened dataset retains the majority of the original dataset's content.

5.2.2 Pandas Library

Pandas is a Python analytical package. Since its founding by Wes McKinney in 2008 as a solution to his demand for a robust and adaptable instrument for quantitative study, pandas have become one of among the most well-known Python libraries. It features a very vibrant member network. The two main Python libraries used by Pandas are NumPy for computational tasks and matplotlib for displaying information. By acting as a layer around these libraries, pandas makes it possible to use fewer lines of code for using several matplotlib and NumPy functions. For example, you can plot a chart in a few lines using pandas'.plot() function, which integrates several matplotlib functions into one procedure.

```
import pandas as pd
```

Prior to pandas, the majority of researchers prepared and munged information using Python before moving the remainder of their operation to a more domain-dependent language, such as R. Pandas provided two fresh categories of objects for information storage: DataFrames, which have a columnar order, and Series, which have a list-like structure, which simplify analytical work and remove the necessity for additional instruments. Although series have significant uses, the majority of analysts deal primarily with information stored in DataFrames. Similar to a workbook or database, data frames contain information in the well-known table structure of rows and columns. Many statistical operations, including determining the averages per column in a dataset, are made easier by DataFrames.

5.2.3 NumPy Library

The core Python library for computational science is called NumPy. A complex array item, different obtained objects (like masked arrays and matrices), and a variety of habits for quick array operations—like sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random training, and much more—are all provided by this Python library. The ndarray entity is the central component of the NumPy collection. This contains uniform kinds of data in n-dimensional arrays, and numerous operations are carried out in compilation for

efficiency. The expression "broadcasting" refers to the implied conduct of processes, item by item. Furthermore, if the lesser array is "expandable" to the form of the greater array in an approach that makes the resultant broadcast clear, a and b in the illustration before could be multiple arrays of identical form, a scalar and an array, or even two arrays with distinct forms.

import numpy as np

Once more, NumPy provides complete object-oriented assistance, beginning with ndarray. As an illustration, the class ndarray has many techniques and properties. Programmers are free to choose whatever paradigm they like to use while coding because a lot of its techniques are replicated by procedures in the outside NumPy namespace.

5.2.4 Matplotlib Library

Matplotlib is a multi-platform toolkit for Python and its numerical module NumPy that facilitates data visualization and graphical charting (bar graphs, scatter diagrams, histograms, etc.). Therefore, it provides a strong open-source substitute for MATLAB.

import matplotlib.pyplot as plt

Diagrams can also be included in GUI programs by programmers using the matplotlib APIs (Application Programming Interfaces). The design of a Python matplotlib script makes it possible to create an interactive data plot with only a few lines of code in most cases. Under the matplotlib coding level, two APIs are covered:

- Matplotlib is at the highest level of the Python code object hierarchy that makes up the pyplot API.Pyplot
- More flexible than Pyplot, its Object-Oriented API collection of objects can be created. Immediate access to Matplotlib's underlying layers is possible with this API.

The pyplot API features a handy state-driven approach modeled after MATLAB. Actually, the matplotlib Python library was created as an open-source MATLAB substitute at first. Although it is thought to be more challenging to use, the OO API and its interface are far more effective and configurable than Pyplot.

5.2.5 SciPy Library

A free to download Python package called SciPy is employed to address mathematical and scientific issues. Based on the NumPy extension, it gives the user access to a vast array of high-level functions for data manipulation and visualization. Since SciPy utilizes NumPy, as was previously explained, importing SciPy eliminates the requirement to import NumPy. Because SciPy is based on NumPy, you can manipulate arrays by using NumPy functions directly.

```
from scipy import stats
```

5.2.6 Seaborn Library

A Python package called Seaborn is used to create statistical visualizations. It strongly interacts with pandas data structures and relies upon the matplotlib framework. Seaborn facilitates data exploration and comprehension. Its charting methods work with dataframes and arrays that hold entire datasets, and they internally carry out the statistical aggregating and semantic mapping required to create visually appealing graphs. You may concentrate on the meaning of the various plot parts rather than the specifics of how to design them thanks to its explicit, dataset-oriented API.

```
import seaborn as sns
```

There isn't a single, optimal method for visualizing data. Various plots work best for responding to various inquiries. With Seaborn's uniform dataset-oriented API, switching among several graphical representations is simple. The mean of a single factor as a function of other factors is frequently of interest to us. The goal of several specific plot types in Seaborn is to visualize information that is categorical.

5.3 Exploratory Data Analysis (EDA)

EDA is a method of characterizing the information using statistical and graphical tools to highlight significant features for additional examination. This entails examining the dataset from several perspectives and providing a description and summary of its contents without assuming anything. Before tackling statistical modeling or machine learning, EDA is an essential procedure to perform to make sure the information is indeed what it seems to be and that there are no evident errors. Many popular EDA libraries are available for Python, such as NumPy, Matplotlib, Seaborn, Plotly, and pandas.

John Tukey advocated for exploratory data analysis in order to inspire statisticians to investigate information and maybe develop theories that could lead to additional data collecting and experimentation. EDA is more specifically concerned with verifying the presumptions needed for testing hypotheses and fitting models. In addition, it handles values that are absent and modifies variables as necessary while checking. EDA creates a solid grasp of the data and problems related to the information or procedure. There are four types of EDA:

- Univariate Non-graphical
- Multivariate Non-graphical
- Univariate graphical
- Multivariate graphical

5.3.1 Univariate Non-graphical: The most basic type of data analysis is called univariate non-graphical because it just uses a single factor to get information. Understanding the fundamental variation in samples and information, as well as drawing conclusions about the population, are the usual objectives of univariate non-graphical EDA. The evaluation also includes the identification of outliers.

Typical or middle values are related to the distribution's position or central trend. The most frequently employed statistics for calculating central patterns are mean, median, and occasionally mode; mean is the most frequently employed. Distribution serves as

a gauge for how far away from the center we are when looking for information components.

Short informative factors known as descriptive statistics are used to provide an overview of a specific data collection, which may be a sample or a representative of the full population. By providing brief descriptions of the collection and data measurements, statistical analyses aid in the description and comprehension of the characteristics of a particular dataset.

Iris_df.describe()

In our implementation we have outputted some statistics regarding the iris dataset. The spread and primary tendency of the numerical parameters in the dataset are briefly summarized by these statistics.

Table 3. The statistical information for Iris dataset

	sepal length (cm)	**sepal width (cm)**	**petal length (cm)**	**petal width (cm)**
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

So, the outcome shows us that in this dataset we do not have values that are absent. We have 150 items for each characteristic of the flowers. Then we have the mean for every characteristic, such as the mean value for sepal length is 5.84333 cm, and so on. We have as well the standard deviation to determine how numbers are distributed or

dispersed around the mean. Then, the minimum value for every category is shown as well, the maximum value in cm for every characteristic is shown in the table.

iris_df.species.describe()

The outcome shows that there are in total 150 values, we have 3 unique type of flowers, the top one is setosa, and the frequency for each type is divided equally with 50 species each. The name is species and the data type are object.

print(life.describe())

We have printed descriptive numerical values for every column in life expectancy dataset for further analysis.

5.3.2 The multivariate non-graphical: This EDA method is typically employed in statistical or cross-tabulation contexts to illustrate the relationship among multiple variables. Cross-tabulation is a very helpful tabular tool for categorical information. We generate statistics for every stage of each categorical variable and one quantitative variable independently, then we evaluate the results over the total number of categorical variables.

5.3.3 Univariate graphical: Although non-graphical approaches are unbiased and numerical, they are unable to provide an accurate representation of the facts; Because they necessitate a certain amount of subjective analysis, pictorial approaches are consequently more frequently used. Typical examples of univariate graphics include:

Histogram: A histogram is the simplest fundamental type of chart; it might be a plot of bars where each bar indicates the frequency (count) or proportion (count/total count) of occurrences for a range of values.

iris_df.hist();

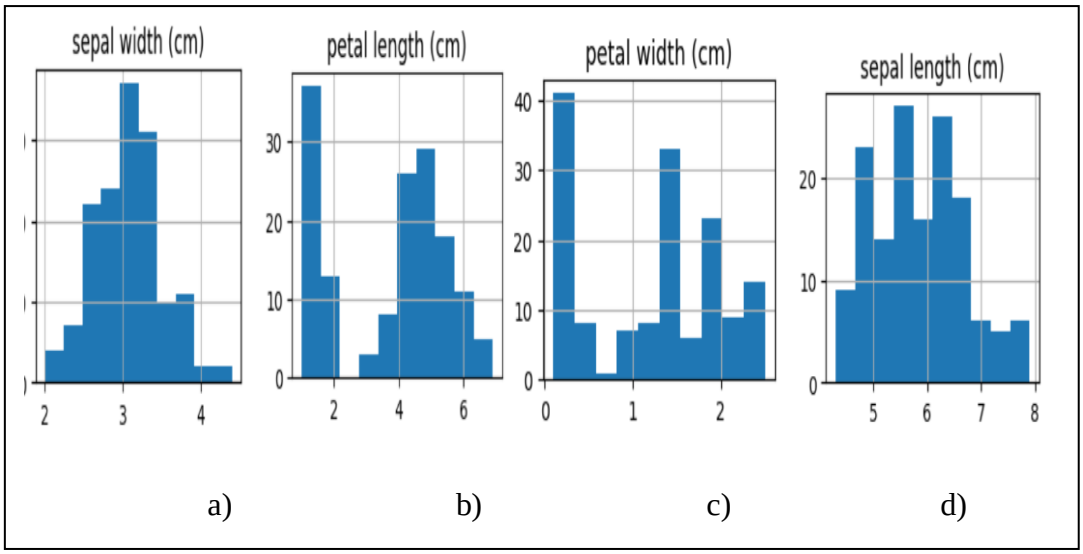

Figure 5.3.3. 1 The distribution of sepal, petal length and width

We use the hist() function in order to see how different number factors are distributed throughout the Iris dataset.

iris_df.hist(column='sepal length (cm)', by='species', figsize=(12, 3), layout=(1, 3));

Here, we have built a histogram where we have specified the column as sepal length (cm) foe every species, we have set the size of the histogram with dimensions 12 and 3 and have set a horizontal layout.

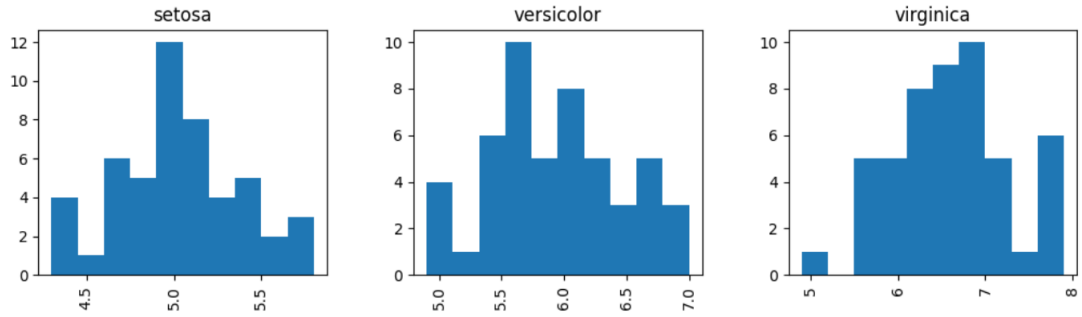

Figure 5.3.3. 2. Distribution among species of Iris Dataset

We have presented another histogram that shows IQ values. We have assigned values to mean (mu) variable and standard deviation (sigma) as 100 and 15 respectively. The value of samples is 10000. We produce random data using the Gaussian dispersion.

 # Parameters for generating random data

mu, sigma = 100, 15

num_samples = 10000

 # Generate random data using NumPy

x = mu + sigma * np.random.randn(num_samples)

 # Plot the histogram of the data

n, bins, patches = plt.hist(x, 50, density=1, facecolor='g', alpha=0.75)

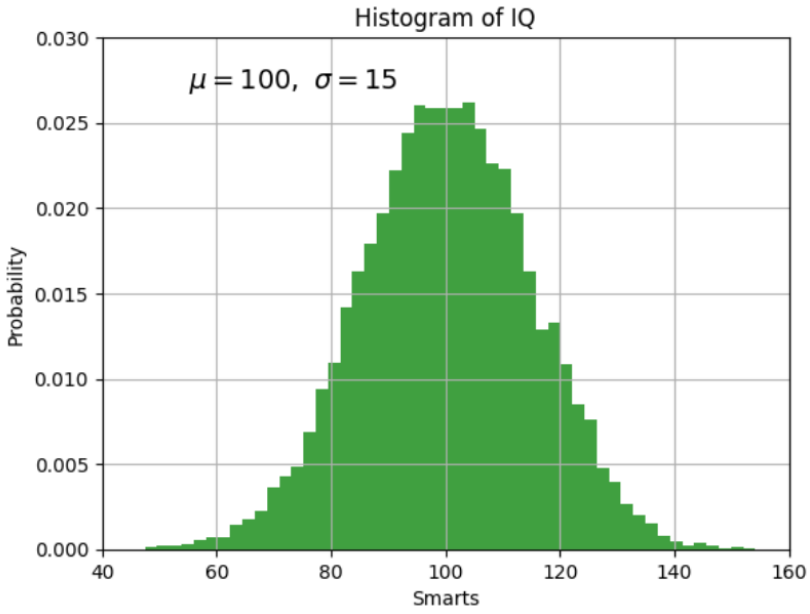

Figure 5.3.3. 3. The representation of numerical values in Life Expectancy Dataset

plt.hist(life['Life_expectancy '], bins=20, color='skyblue', edgecolor=black'

We have represented the numerical values in the descriptive analysis of the dataset Life Expectancy in a histogram. We have assigned the intervals as 20, the color of the histogram as sky blue and the color of the edges of the table as black.

Stem-and-leaf plots: These can be a simple alternative to a histogram. It displays every data element and, consequently, the distribution's form.

Boxplots: The boxplot is a very helpful univariate visual approach. Even though they will be deceptive regarding elements like multimodality, boxplots are excellent at

showing data on central tendency, accurate indicators of place and spread, as well as offering data about symmetry and anomalies.

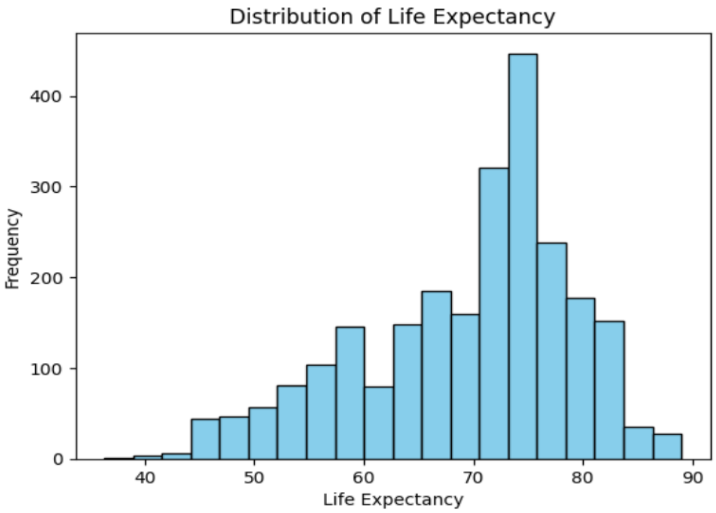

Figure 5.3.3. 4. Distribution of Life Expectancy dataset

We use the function boxplot() from library pandas in order to see the species-specific variation in sepal length.

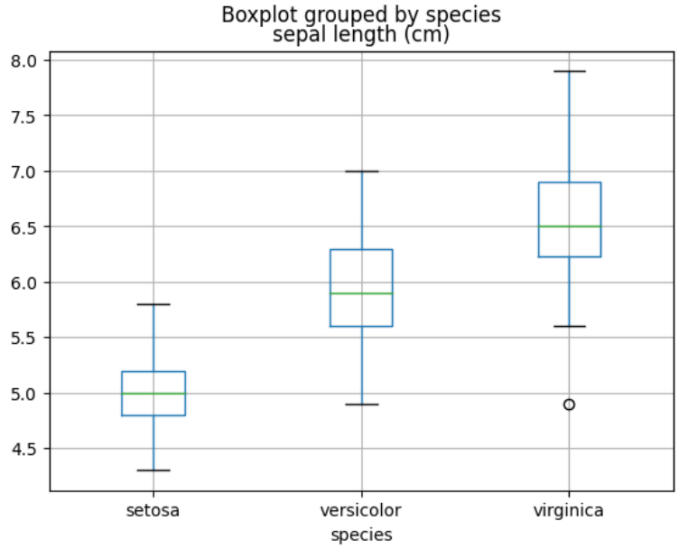

Figure 5.3.3. 5. Species specific variation in sepal length

The upper – for every species shows the maximum value, the upper line of the rectangle for each representation is the upper quartile, meaning that every value greater

than that of the line belongs to the 25% of values higher than it. The lower line of the rectangle for each representation is the lower quartile, meaning that every value below the value of the line belongs to the 25% of values lower than it. The middle line of the rectangle is the median, meaning that values greater than its value represent the 50% of values greater than its value and those lower than it represents the 50% of values lower than its value. The lower – for every species shows the minimum value in each category.

Plots with quantile normality: The most complex univariate graphical EDA method. It's common to check how closely a given sample adheres to a certain hypothetical distribution.

5.3.4 Multivariate graphical data: This type of data shows the connections among multiple collections of information through the use of visuals. The only one that is frequently employed is probably an organized bar plot, where every set represents the value of one of the parameters and each bar in a gaggle represents the quantity of the other variable.

Additional typical types of multivariate visuals include:

Scatterplot: The scatterplot, which has a single value on the x-axes and y-axes and, consequently, the point for every scenario in the dataset, is the fundamental visual EDA approach for two numerical variables.

Run chart: It is an information line chart with time intervals shown.

Heat map: It is a visual format for information where colors are used for representing values.

5.4 Linear Regression

Determining the linear connection among a goal and one or more variables is done using linear regression. Basic and complex linear regression are the two varieties.

When determining the association among two continuous variables, simple linear regression is helpful. There are two types of variables: independent or forecast and dependent or response. Statistical relationships are sought after rather than predictable ones. When two distinct factors can be precisely defined by one another, a relationship is said to be predictable. Finding the line which most matches the information is the main concept. The segment with the lowest overall forecasting error (across all of the information values) is the most accurate match. The gap among the point and the regression line is called the error.

$$Y(pred) = b0 + b1*x \quad (1)$$

It is necessary to select b0 and b1 variables in a way that minimizes mistake. The objective is to find the path that minimizes the deviation the most if the sum of squared errors is used as the model's evaluation measure.

The beneficial and detrimental points will cancel one another out if the mistake is not squared. Regarding a single predictor model:

Investigating "b1":

- The association between x(predictor) and y(target) is positive if b1 > 0. In other words, a rise in x will raise y.
- The association between x(predictor) and y(target) is negative if b1 < 0. Thus, a rise in x will result in a fall in y.

Examining "b0":

- The forecast using just b0 in Equation 1, will be worthless if the model does not contain x=0. As an illustration, we are using a dataset that links weight (y) and height (x). Assuming that height is zero (x=0), the formula will have merely a value of b0, which is nonsensical because weight and height cannot possibly be zero in real life. Considering the algorithm's data outside of its parameters led to this outcome.

- "b0" is the mean of all expected outcomes when x=0 if the model has value 0. However, it is frequently hard to set all of the factors that predict to zero.

- The remainder has an average of zero if b0 is present. Regression will be required to pass over the origin if there isn't a "b0" term. There is going to be bias in the forecast as well as the regression co-efficient.

In addition to the formula before, the standard formula can also be used to obtain the algorithm's coefficient.

The co-efficient of each indicator, including the fixed term "b0," is contained in theta. The standard formula computes by obtaining the input matrix's reverse. As the number of characteristics increases, so will the computation's difficulty. As the number of characteristics increases, it becomes extremely sluggish. Because of its complexity, the standard formula is challenging to apply; this is where the technique of gradient descent is useful. The ideal coefficient value can be obtained by taking an incomplete derivative of the cost function in relation to the variable.

Residual Analysis: In a regression framework, unpredictability and randomization are the two main constituents.

Prediction = Deterministic + Statistic

The predictor variable in the model covers the predictable portion. The notion that both the actual and projected values are uncertain is revealed by the stochastic portion. There will continue to have some details that are overlooked. The leftover data contains this data. Utilizing the results of the remains, the remainder plot facilitates model analysis. Plotting is done among residue and anticipated values. They have uniform values. The point's separation from 0 indicates how inaccurate the forecast was for that particular number. The forecast is poor if the value is positive. The forecast is high if the outcome is negative. A number of zero denotes a perfect forecast. Finding residual patterns can help the model get better.

R-Squared value: This number is between 0 and 1. A value of "1" means that the predictor fully explains all of the variation in Y. Predictor "x" has a value of "0," meaning that it does not explain any variance in "y."

Sum of Squared error (SSE): The range of the goal number (estimated result) along the regression line.

Low P-value: Dismisses the null assumption, which suggests a relationship between the outcome and the expected value.

High P-value: Targeted adjustments are not correlated with predictor adjustments.

CHAPTER 6

EXPERIMENT AND RESULTS

We have organized this part into three study cases. The first study case performs some statistical tests, data analysis on the datasets utilized in the methodologies. We have also analyzed the correlation matrix to identify the relationship among the variables. The second and third study case consist on answering the research questions raised in the hypothesis. The second study case uses the Iris dataset, while the third uses Life Expectancy dataset.

6.1 First Study Case

We analyzed the correlation matrix for both frameworks, for the same dataset: Life Expectancy dataset. A correlation matrix can be used to reduce a substantial quantity of information, recognize patterns, and make decision relating to it. One statistical method for assessing the connection among two variables in a data set is to create a correlation matrix. The matrix is a chart where each cell has a correlation coefficient, with 1 denoting a strong association, 0 a neutral relationship, and -1 a weak relationship between the variables.

We need to check the correlations among variables in the dataset, since a high number of correlations in linear regression indicates that the results will not be trustworthy.

The values generated were the same, as we can notice on both the graphs below.

We can distinguish some high values of correlations in this matrix. Life expectancy variable and Income composition of resources have a correlation coefficient of 0.732734, that shows that nations with longer life expectancies typically have more revenue distributed among their resources. Another example of a high correlation coefficient value is that among infant deaths and under five deaths variables. The value is 0.996629 and implies that shows these two factors have a strong correlation, which

is apparent given that baby deaths are included in the category of mortality for children under five. We can clearly observe from the matrix the values that have high correlations due to stronger colors of red and blue, and we can imply that there are very few high values of correlation, resulting in reliable outcomes of the linear regression model.

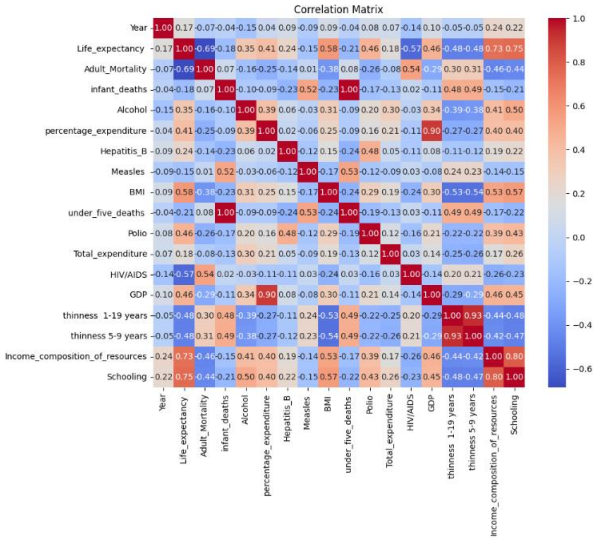

Figure 6.1. 1. Correlation Matrix for Life Expectancy dataset in OSEMN methodology

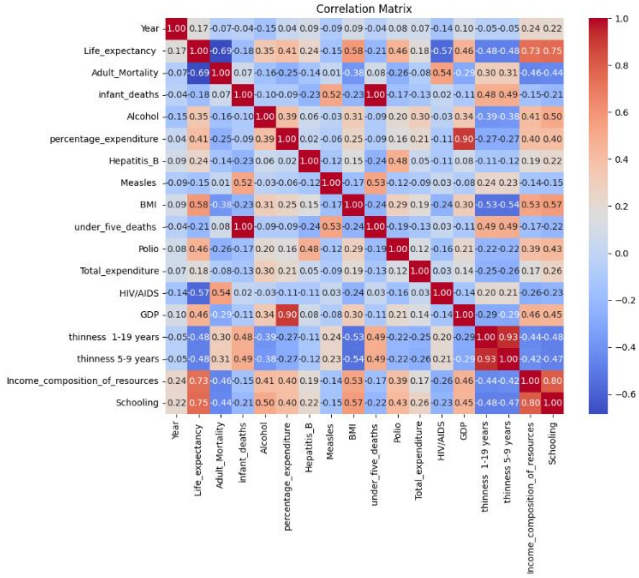

Figure 6.1. 2. Correlation Matrix for Life Expectancy dataset in CRISP-DM methodology

 # Perform Pearson correlation between variables

```
r, p = stats.pearsonr(life['GDP'], life['Life_expectancy '])
print("Pearson correlation between GDP and Life Expectancy: r={:.3}, p-value={:.3}".format(r, p))
```

We have displayed the Pearson correlation among variables GDP and Life Expectancy in the CRISP-DM framework. The outcome values are: r=0.46 and p-value=3.55e-129.

```
from scipy import stats
```

 # Pearson correlation between GDP and Life Expectancy

```
r, p = stats.pearsonr(life['GDP'], life['Life_expectancy '])
print("Pearson correlation (GDP vs. Life Expectancy):\n")
print("Correlation coefficient (r): {:.3f}".format(r))
print(" p-value={:.3}".format(p))
```

We have displayed the Pearson correlation among variables GDP and Life Expectancy in the OSEMN framework. The outcome values are: r=0.46 and p-value=3.55e-129.

The r and p-values for both frameworks are the same.

Value r=0.46 warrants a moderate positive progression of the GDP with the life expectancy. As the GDP rises, the life expectancy also rises and vice versa implying that there is a positive relationship between GDP and life expectancy of a given country. The points closer to the upper right area of the plot represent better positive linear correlation or r closer to 1.

This is an insignificantly small p-value and according to statistical practice, a small p-value further substantiates the rejection of the null hypothesis. Thus, the null hypothesis we shall use for comparison would be that there is no relationship between GDP and life expectancy. Since p-value is less than traditional levels of significance, it is possible to infer that there is a significant relationship between levels of GDP and life expectancy.

We have performed t-test between sepal width of setosa and virginica species for both frameworks.

```
t, p = stats.ttest_ind(iris_df[iris_df.species =='virginica'] ['sepal width (cm)'],iris_df[iris_df.species =='setosa'] ['sepal width (cm)'])
print ("t-score = {:.3}, p-value= {:.3}".format(t, p))
```

For OSEMN framework, the outcomes from the pseudocode displayed above are:

t-score = -6.45, p-value= 4.25e-09.

A t-score of -6.45 informs us that the virginica species has a shorter mean sepal width as compared to setosa species since the t-score is a negative value. The obtained t-score shows that there is a significant difference in the means of the two groups.

The low p-value that is in the range of 0 and indicates that there is only a 0.01% chance it could be due to random chance indicates that there is strong evidence against the null hypothesis. Hence, we can decisively discard the null hypothesis and conclude that indeed the means for sepal width of Setosa and Virginica species are significantly different.

```
# Perform statistical tests
# Example: Perform t-test between sepal width of setosa and virginica species
t, p = stats.ttest_ind(iris_df[iris_df['target'] == 0]['sepal width (cm)'], iris_df[iris_df['target'] == 2]['sepal width (cm)'])
print("T-test between Sepal Width of Setosa and Virginica species:")
print("t-score = {:.3}, p-value = {:.3}".format(t, p))
```

For CRISP-DM methodology, the values from conducting the t-test are:

t-score = 6.45, p-value = 4.25e-09.

This t-value of 6.45 presents the extent of the variations in means of sepal widths for the Setosa and Virginica varieties of the flowers. The measure of the absolute t-score

implies that the bigger the value of t-score, then the greater the difference in means between the two groups.

The p-value is the same value for both frameworks, and the interpretations remain the same.

```
import scipy.stats as stats

setosa_width = iris_df[iris_df.species == 'setosa']['sepal width (cm)']
versicolor_width = iris_df[iris_df.species == 'versicolor']['sepal width (cm)']
virginica_width = iris_df[iris_df.species == 'virginica']['sepal width (cm)']

f, p = stats.f_oneway(setosa_width, versicolor_width, virginica_width)
print("F-value = {:.2f}, p-value = {:.3}".format(f, p))
```

We have performed ANOVA for sepal width across different species, for both frameworks. Above we have the pseudocode performed for OSEMN methodology. The outcomes of values are: F-value = 49.16, p-value = 4.49e-17.

```
# Perform ANOVA
# Example: Perform ANOVA for sepal width across different species
f, p = stats.f_oneway(iris_df[iris_df['target'] == 0]['sepal width (cm)'],
                      iris_df[iris_df['target'] == 1]['sepal width (cm)'],
                      iris_df[iris_df['target'] == 2]['sepal width (cm)'])
print("ANOVA F-value = {:.2f}, p-value = {:.3}".format(f, p))
```

We have performed ANOVA for CRISP-DM framework as well. The outputs of this pseudocode are: ANOVA F-value = 49.16, p-value = 4.49e-17.

We can clearly observe that F-value and p-value for both methodologies have the same values. This is known as the F statistic of ANOVA test or simply known as F test. It indicates the amount of variation or dispersion across the groups and that within the groups. A larger F-value indicates that the variability between the groups is more than

the variance established within each of the groups. The F-value of 49.16 shows that there is a significant difference between the group means.

Giving data to the test, the respective p-value is equal to 4. 49e-17 (or approximately 0). This very low p-value means that there is enough evidence at very high significance to reject the null hypothesis thus explaining why there is significant variation across the groups.

From this study case we conclude that since the correlation coefficient values for both frameworks are the same; r, p-value, F-score, t-score values used in conducting a statistical analysis, are the same for both methodologies.

This implies that the statistical evidence arrived at is consistent whichever model is used to analyze them. It amplifies the reliability and confirms that the relationship between the variables is not affected by any specific method of computation used in the study between the GDP and life expectancy.

6.2 Second Study Case

In this second study case we have utilized the Iris dataset. We will initially examine the model part of CRISP-DM methodology. We have presented the procedure of how to train as well as test a linear regression model with the aim of making a prediction of sepal width from the other variables included in the iris data set.

 # Split the data into training and testing sets

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

First, by utilizing the help of train_test_split function belonging to sklearn library of python, it divides the dataset into a training dataset and a testing dataset. The model selection assists in checking the behavior of the model when exposed to new datasets which have not been used in the training process.

 # Train a linear regression model

```
model = LinearRegression()
model.fit(X_train, y_train)
```

 # Predict sepal width for the testing set

```
y_pred = model.predict(X_test)
```

It builds a linear model (LinearRegression) that is based on the training set. The model will be trained for classifying the target parameter ‘sepal width (sepal width (cm))’ out from the data set ‘X’. On the basis of the model fitted with the given data, it gives predictions (y_pred) for sepal width using the testing dataset (X_test).

 # Calculate Mean Squared Error (MSE) and R-squared

```
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print("Mean Squared Error (MSE):", mse)
print("R-squared:", r2)
```

Mean Squared Error measures the average of the squared differences between the prediction and the actual values, on the other hand R-squared measures how efficiently the values of the target variable can be predicted using the values of independent variables.

 # Visualize the results: Predicted vs Actual Sepal Width

```
plt.figure(figsize=(8, 6))
plt.scatter(y_test, y_pred, color='skyblue', alpha=0.7)  # Scatter plot
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], linestyle='--', color='red')  # Diagonal line
plt.title("Predicted vs Actual Sepal Width")
plt.xlabel("Actual Sepal Width")
plt.ylabel("Predicted Sepal Width")
plt.grid(True)  # Add grid lines for better readability
plt.show()
```

We have visualized the plotting of the actual sepal width (y_test) against the predicted sepal width (y_pred) with a scatter plot. The straight line on the left of the figure signifies the ideal condition where the observed values are fully predicted by the model at hand and the proximity of the points to this line indicates the level of accuracy of the model.

The outcomes from the pseudocode are the Mean Squared Error (MSE) and R-squared, as well as the scatter plot where we can observe the relationship between the actual values and predicted ones.

The error rate of the estimator or forecasting model developed using the provided set of sample data is represented by the mean squared error, or MSE. In order to calculate the disparity among the predictions made by the model and real inquiries, it calculates the average squared difference among the predicted and actual values. The regression model is more accurate and its predictive accuracy is higher when the mean square error (MSE) is smaller.

Mean Squared Error (MSE): 0.00
R-squared: 1.0

As we can observe from the output, the MSE value is relatively small and extremely close to 0. This means that the sum of the squares of the individual differences in sepal width predictions and the actual sepal widths for the different types of species is almost zero. This implies that the model's forecast is very accurate in relation to the actual values thus making it suitable to be used in estimating values based on given model. The predicted values are very close to the real ones.

In a regression model, R-Squared (also known as R2 or the coefficient of determination) is a statistical metric that establishes how much of the variance in the dependent variable can be accounted for by the independent variable. So, the goodness of fit, or r-squared, indicates how well the data fit the regression model.

As we can observe the value of r-squared is 1, meaning a perfect model fit. When a model's fit to the data is perfect, as shown by a value of 1.0, all variance in the target variable is explained by the model.

As a conclusion, analyzing the information shown in the table and the graphics the extremely low MSE, as well as the perfect R-squared value are pointing to the fact that the linear regression model is working perfectly on the testing data providing the minimum error. This indicates that for features extracted from the dataset, the model correctly identifies sepal width.

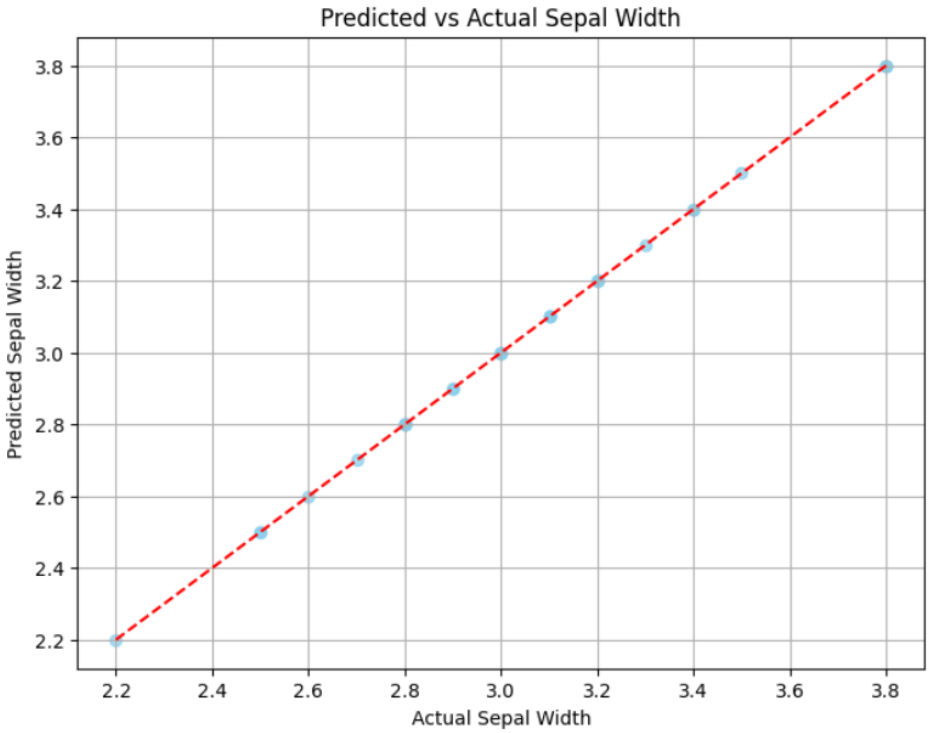

Figure 6.2. 1. Predicted vs Actual Sepal Width Values presented in a graph at the modeling part of CRISP-DM methodology

As we can observe from the graph, the predicted and actual values lie in the same line, indicating that the model have predicted accurately the values. This analysis is followed by a residual analysis, which indicates that the linear regression model that has been built based on the CRISP-DM methodology is indeed very accurate and the model fits the data so well.

We have examined the part of modeling at the OSEMN methodology as well. We have applied the linear regression and used in the prediction of sepal width given that the species is Virginica species and assesses the validity of the model based on the MSE and the R Squared scores.

```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Train a linear regression model
model = LinearRegression()
model.fit(X_train, y_train)
# Predict sepal width for the testing set
y_pred = model.predict(X_test)
```

The dataset is divided into the variables or independently for inputs (X) and one as the dependent or output variable (y). It has columns like; Species which is later coded to binary known as is_virginica, to illustrate the existence of species virginica. These sets are again divided into training set and testing set using 'train_test_split' function from sklearn. model_selection.

Sklearn has been employed to instantiate a linear regression model through a LinearRegression() function. linear_model. In the training section, the fit method is used to fit the model on the training data where it locates the associations between all the features and the target outcome.

The trained model is used to compute the sepal width in the testing set or the prediction of outcome of sepal width (X_test). They are saved in the y_pred variable.

```
# Calculate Mean Squared Error (MSE) as a performance metric
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print("Mean Squared Error (MSE):", mse)
print("R-squared:", r2)
```

MSE and the coefficient of determination R-squared scores are used to assess the model's accuracy. MSE stands for mean squared error and it quantifies the average squared deviation of the sepal widths from the predicted sepal widths. Adjusted R-squared is the measure of the model's overall goodness of fit by showing the proportion of the variance in the target variable that is predictable from the set of independent variables. The above metrics are displayed in console.

```
plt.scatter(X_test['is_virginica'], y_test, color='blue', label='Actual')
plt.plot(X_test['is_virginica'], y_pred, color='red', label='Predicted')
plt.xlabel('Presence of Virginica Species')
plt.ylabel('Sepal Width (cm)')
plt.legend()
plt.title('Linear Regression Prediction')
plt.show()
```

The script plots the actual sepal width (y_test) against its prediction (y_pred) in order to produce a scatter chart. In this graph, the dependant variable, sepal width, is marked on the y-axis while the independent variable, the species (Virginica) is marked on the x-axis. The scatter plot also consists of a line which exhibits where the linear regression is predicting the data to be.

The outcomes from the pseudocode are the Mean Squared Error (MSE) and R-squared, as well as the scatter plot where we can observe the relationship between the actual values and predicted ones.

Mean Squared Error (MSE): 0.14378783721993596
R-squared: -0.0050408006985274145

A MSE value of 0.1438 approximately represents the average of the squared differences between the predicted and the actual values. Specifically, the MSE means that, on average, the squared deviation between forecast and actual values of the target variable (sepal width).

The R-square is the statistical measure showing the degree of variance of the target variable accounted for by the independent variables. If the value of R-squared is negative that means that the given model is not better than the horizontal line. Value is -0.00504 approximately. The negative R-squared therefore depicts that even though the linear regression model has been fitted the target variable DON, then it does not adequately explain much of the variance in the variable and hence the model may not be the best model for the data.

These values give some evidence on how well the linear regression model that was developed using the OSEMN method will perform. The MSE shows the scope of the errors in the predictions and we can observe that the coefficient indicates a considerable amount of errors, while the negative value of R-squared indicates poor fit of the model in explaining the data.

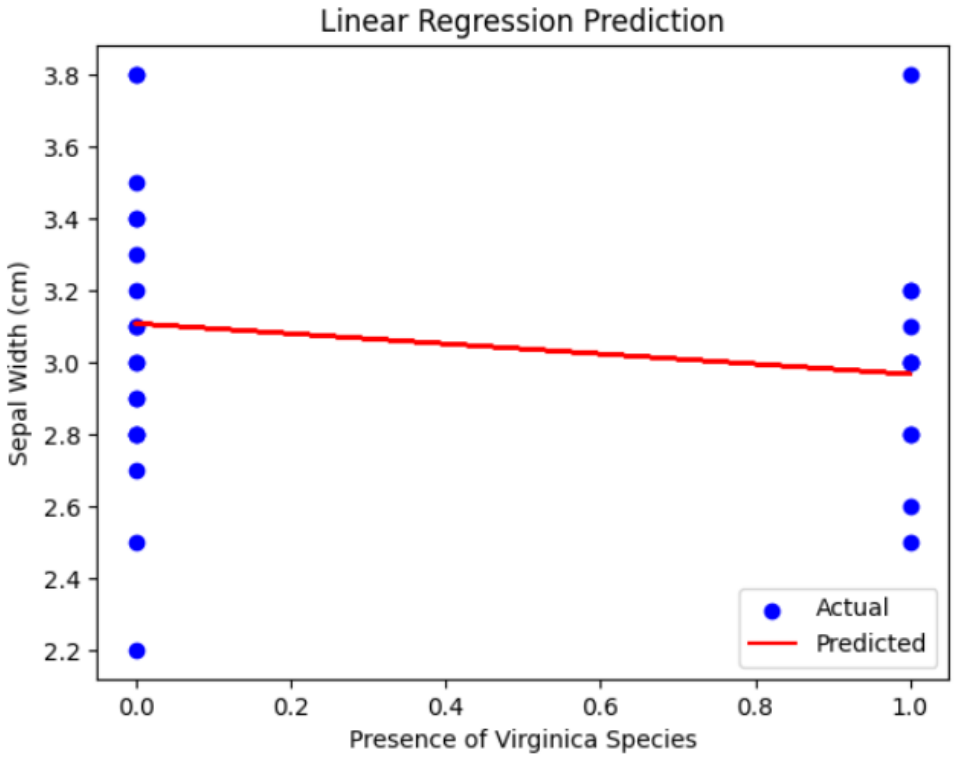

Figure 6.2. 2. Predicted vs Actual Sepal Width Values presented in a graph at the modeling part of OSEMN methodology

As we can observe from the graph, it is very clear the disproportionality among actual and predicted values. Actual values are very different from those predicted indicating

not only a low prediction rate, but also a low model fit and not accurate predicted values.

When comparing values generated from both methodologies, we can suggest the following observations:

- For the CRISP-DM, the MSE value of the model was nearly 0, which reflected the model's nearly perfect accuracy, while for the OSEMN, the value of MSE was around 0. 144, which is substantially higher than one, meaning a greater prediction error.

- This means that the model developed using CRISP-DM methodology is much more accurate in predicting sepal's width than the model developed using the OSEMN methodology.

- Whereas the R-squared value for the OSEMN approach was roughly -0.005, suggesting a very poor fit of the model, the R-squared result for the CRISP-DM methodology was 1.0, showing an ideal fit of the model to the data.

- The obtained negative R-squared value indicates that the OSEMN methodology has constructed a model that is even less accurate than the linear model with the intercept equal to one that simply oscillates around the horizontal line, not managing to unveil connections between the features array and the target variable.

6.3 Third Study Case

In this third study case we have utilized the Life Expectancy dataset. We will initially examine the model part of CRISP-DM methodology. We have presented the procedure of how to train as well as test a linear regression model with the aim of making a prediction of variable "Life expectancy" by utilizing other variables from the dataset: GDP, schooling, alcohol, adult mortality. Multiple linear regression is used

because the independent variables or predictor variables therefore exposed are more than one and can be used to predict the dependent variable which is life expectancy.

This essentially entails identifying the business need or requirement for the upcoming changes phase. In this pseudocode, the aim is to create a prediction model for life expectancy relative to some features, such as GDP, Schooling, and Alcohol consumption. The data has also undergone preprocessing and cleaning where any missing values were removed (using dropna() function) as well as feature selection was also conducted in order arrive at the most relevant features from the large amount of data available.

```
# Split the data into features (X) and target (y)
X = life[features[:-1]] # Features: all columns except the last one (Life expectancy)
y = life[features[-1]]  # Target: Life expectancy
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

We split the information into training and testing sets. We as well separate the target variable that will be predicted from other variables of the dataset.

```
# Modeling
# Train a linear regression model
model = LinearRegression()
model.fit(X_train, y_train)
# Predict life expectancy for the testing set
y_pred = model.predict(X_test)
```

Based on the selected features, the model is fitted on the training data (X_train, y_train) using the LinearRegression class in scikit-learn. Then the variables for testing are used in the model and the results are being stored in the y_pred variable.

```
# Evaluation
# Calculate Mean Squared Error (MSE) and R-squared
```

```
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print("Mean Squared Error (MSE):", mse)
print("R-squared:", r2)
```

This gives the mean of the squared differences between the actual life expectancy values taken from the test dataset (y_test) and the predicted life expectancy values obtained from the chosen model (y_pred). MSE aims at calculating the mean of squared deviations of fitted values from the observed values of the dependent variable; it thus gives a measure of how close the predicted values are to the actual values in the sample set. Low value of MSE indicated a better accuracy.

This deploys the formula that determines the coefficient of determination, also known as R-squared, by comparing the actual life expectancy values (y_test) and the predicted life expectancy values (y_pred). R-squared measures the percentage of variance in dependent variable (life expectancy) that characterized by independent variables/features in the model. R-squared values have a range from 0 to 1, and values closer to 1 indicate a better model fit.

```
# Visualize the results: Predicted vs Actual Life Expectancy
plt.figure(figsize=(10, 6))
plt.scatter(y_test, y_pred, color='skyblue', alpha=0.7)  # Scatter plot
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], linestyle='--', color='red')  # Diagonal line
plt.title("Predicted vs Actual Life Expectancy")
plt.xlabel("Actual Life Expectancy")
plt.ylabel("Predicted Life Expectancy")
plt.grid(True)  # Add grid lines for better readability
plt.show()
```

These lines of code create a visual representation of the predicted values as well as the actual ones. We have set the size of the figure, of the scatter plot, its representing colors, and we also assed grid lines to ensure better readability.

The outcomes from the model are the Mean Squared Error (MSE) and R-squared, as well as the scatter plot where we can observe the relationship between the actual values and predicted ones.

Mean Squared Error (MSE): 35.86476496218019
R-squared: 0.43992323085617346

A MSE value of 35.8648 approximately represents the average of the squared differences between the predicted and the actual values. Specifically, the MSE means that, on average, the squared deviation between forecast and actual values of the target variable (sepal width).

The square error where the predicted life expectancy values as calculated from the testing set is having a value of 35.86476. A lower MSE value is preferred, albeit by a small margin, as it implies a smaller difference between the model's prediction and the actual value designated during feature extraction and model training. This value shows that there are errors among the predicted values and actual ones.

The R-square is the statistical measure showing the degree of variance of the target variable accounted for by the independent variables. If the value of R-squared is negative that means that the given model is not better than the horizontal line. The R-squared value of 0.4399 approximately indicates that such features could account for about 44% of the total variation in life expectancy to the models. This shows a moderate degree of explanatory fits. As the model can explain a significant portion of variation in life expectancy, there is still the major part of variation unexplained.

As we can observe from the graph, we can clearly distinguish that the actual values are very close to the predicted line. Few values are far from the predicted line, but this can be associated to exceptions in the relationship between variables, since some variables van have non-linear relationship with one another. However, even when a perfect model is achieved, and all measurements are accurate, there will always be that component of randomness in the data. They simply fluctuate since randomness is innate to such data; some points could shift from the line that had been expected. Overall, the graph presents a good model fit, and accurate prediction of values.

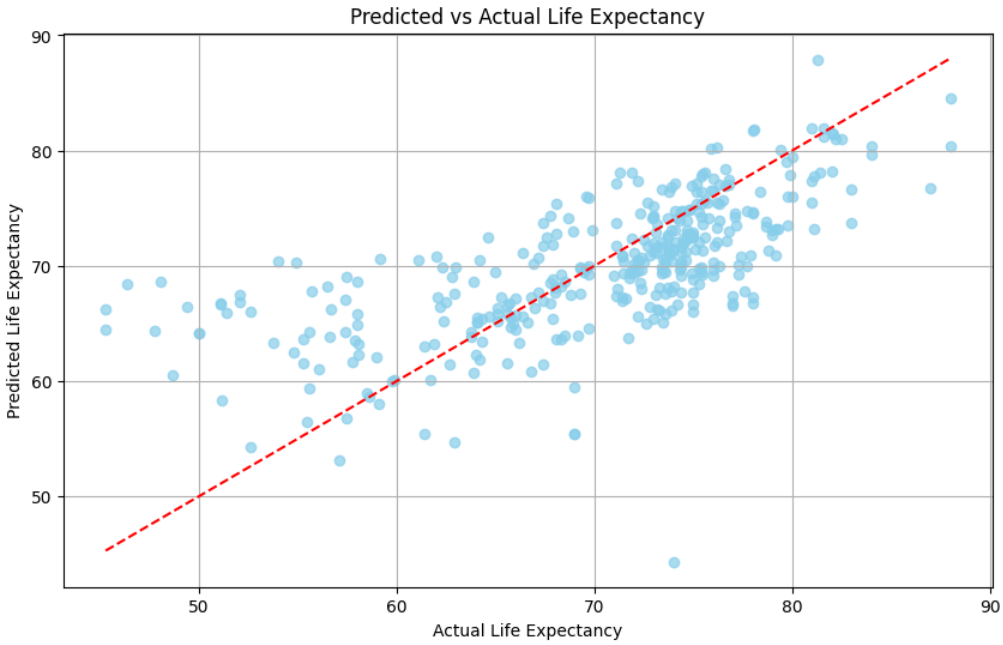

Figure 6.3. 1. Predicted vs Actual Life Expectancy Values presented in a graph at the modeling part of CRISP-DM methodology

We will examine the model part of OSEMN methodology. We have presented the procedure of how to train as well as test a linear regression model with the aim of making a prediction of variable "Life expectancy" by utilizing other variables from the dataset: GDP, schooling, alcohol, adult mortality. Multiple linear regression is used because the independent variables or predictor variables therefore exposed are more than one and can be used to predict the dependent variable which is life expectancy.

```
from sklearn.impute import SimpleImputer
# Initialize the imputer
imputer = SimpleImputer(strategy='mean')
# Fit and transform the imputer on the training data
X_train_imputed = imputer.fit_transform(X_train)
# Transform the test data using the fitted imputer
X_test_imputed = imputer.transform(X_test)
# Train a linear regression model
model = LinearRegression()
model.fit(X_train_imputed, y_train
```

 # Predict life expectancy for the testing set

y_pred = model.predict(X_test_imputed)

It uses SimpleImputer from the Scikit-Learn library and the imputation method as 'mean'. The 'mean' strategy imputes the missing data in the dataset through replacing those data with the mean of the non-missing data in that particular column. It uses the imputed training data to train a linear regression model and then employs the model to forecast the life expectancy of the testing set.

 # Transform the test data using the fitted imputer

X_test_imputed = imputer.transform(X_test)

 # Predict life expectancy for the testing set

y_pred = model.predict(X_test_imputed)

This code provides that the data being tested is handled in the same manner as the training set (i.e., the mean is used to estimate missing values), and it then applies the trained model to the processed test data to provide predictions.

from sklearn.metrics import mean_squared_error, r2_score

 # Calculate Mean Squared Error (MSE) and R-squared

mse = mean_squared_error(y_test, y_pred)

r2 = r2_score(y_test, y_pred)

 # Print the evaluation results

print("Mean Squared Error (MSE):", mse)

print("R-squared:", r2)

This gives the mean of the squared differences between the actual life expectancy values taken from the test dataset (y_test) and the predicted life expectancy values obtained from the chosen model (y_pred). MSE aims at calculating the mean of squared deviations of fitted values from the observed values of the dependent variable; it thus gives a measure of how close the predicted values are to the actual values in the sample set. R-squared measures the percentage of variance in dependent variable (life expectancy) that characterized by independent variables/features in the model. R-

squared values have a range from 0 to 1, and values closer to 1 indicate a better model fit.

According to the actual life expectancy numbers, it arranges the indexes of the testing set (y_test) in ascending order. By doing this, the expected and actual values are guaranteed to be correctly aligned for charting. It enhances the plot with a legend that shows the color coding of the expected and actual life expectancy numbers.

The outcomes from the model are the Mean Squared Error (MSE) and R-squared, as well as the scatter plot where we can observe the relationship between the actual values and predicted ones.

Mean Squared Error (MSE): 0.14378783721993596
R-squared: -0.0050408006985274145

A MSE value of 0.1438 approximately represents the average of the squared differences between the predicted and the actual values. Specifically, the MSE means that, on average, the squared deviation between forecast and actual values of the target variable (sepal width).

The R-square is the statistical measure showing the degree of variance of the target variable accounted for by the independent variables. If the value of R-squared is negative that means that the given model is not better than the horizontal line. Value is -0.00504 approximately. The negative R-squared therefore depicts that even though the linear regression model has been fitted the target variable DON, then it does not adequately explain much of the variance in the variable and hence the model may not be the best model for the data.

These values give some evidence on how well the linear regression model that was developed using the OSEMN method will perform. The MSE shows the scope of the errors in the predictions and we can observe that the coefficient indicates a considerable amount of errors, while the negative value of R-squared indicates poor fit of the model in explaining the data.

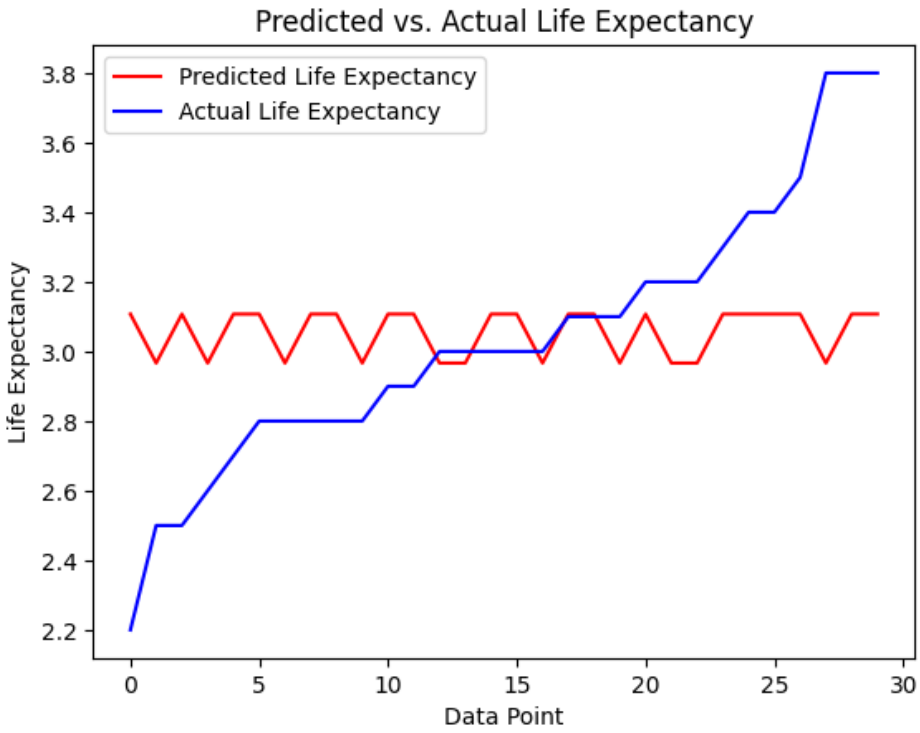

Figure 6.3. 2. Predicted vs Actual Life Expectancy Values presented in a graph at the modeling part of OSEMN methodology

As we can observe from the graph, we can distinguish that the predicted values of life expectancy variable are very different from the actual ones. There are very few values predicted correctly or that have a near distance from one another. This suggests that the accuracy of the predicted values is low and that it has a low model fit.

When comparing values generated from both methodologies, we can suggest the following observations:

- The model that was created using OSEMN methodology shown lower MSE, which means better performance and higher accuracy in the life expectancy prediction in comparison with the model created using CRISP-DM methodology.

- In comparison to the model created using the OSEMN approach, the model created using the CRISP-DM technique has a higher R-squared value, indicating a better overall fit to the data.

- From the graph, we can suggest that the actual values were very close to the predicted values line in the CRISP-DM methodology model, while the actual values were not close to the predicted values line in the OSEMN methodology, indicating that the model of CRISP-DM has in overall a better model fit.

CHAPTER 7

DISCUSSIONS

Only 5 of the 41 papers that were evaluated for full-text relevance during the screening process proved adequate for our study. We have examined the appropriate studies based on the following criteria in an effort to address the research issues listed in the introduction section: the metrics used in determining the model performance; the initial steps performed in the dataset to prepare and handle errors in the data provided.

Methods and standard procedures are not the same thing and while guided processes consist of following a set of stages sequentially, many methods do not. There are two very illustrative situations that are related to data science [1]. The first example is software engineering, which offers a wide range of approaches. Depending on numerous internal and external aspects, none of these approaches appears to be the best option in every circumstance. Although the framework of software development, including that of numerous other engineering issues, is similar to CRISP-DM and OSEMN in many respects (beginning with company requirements and concluding with the deployment and maintenance of the procedure's output), it would also be inappropriate to apply the same linear flow to every problem and scenario.

Science technique is the second example. Instead of being data- or goal-driven, the entire process of scientific discovery is typically question-driven, though it is typically far more adaptable in the early stages (unexpected findings, randomness, etc.). Although some paths in data science bear similarities to scientific approaches, there is a debate about whether data science methodologies should take additional guidance from the general scientific approach or whether data science has replaced the traditional scientific method. While equity and confidentiality were always challenging with data mining, the trajectory model does not yet clearly address all of the legal and ethical concerns surrounding data science, a subject that is becoming more important in data science than the prior data mining paradigm.

Future research aims to achieve transparency at every level of the procedure, making it simple to identify problems and quickly return to a particular stage when necessary for both frameworks.

CHAPTER 8

CONCLUSION

We have answered our research question by considering the results from the experiment part of our paper. To answer the research questions, we can consider the MSE and R-Squared values from the second and third study case. In the second study case, for the CRISP-DM, the MSE value of the model was nearly 0, which reflected the model's nearly perfect accuracy, while for the OSEMN, the value of MSE was around 0. 144, which is substantially higher than one, meaning a greater prediction error. This means that the model developed using CRISP-DM methodology is much more accurate in predicting sepal's width than the model developed using the OSEMN methodology. Whereas the R-squared value for the OSEMN approach was roughly -0.005, suggesting a very poor fit of the model, the R-squared result for the CRISP-DM methodology was 1.0, showing an ideal fit of the model to the data.

While in the third study case, the finding changes slightly. The model that was created using OSEMN methodology shown lower MSE, which means better performance and higher accuracy in the life expectancy prediction in comparison with the model created using CRISP-DM methodology. In comparison to the model created using the OSEMN approach, the model created using the CRISP-DM technique has a higher R-squared value, indicating a better overall fit to the data.

The results of this study support the H0 Hypothesis demonstrating that the CRISP-DM framework has better model fit than OSEMN framework.

To answer the second research question observing the graph, we can suggest that the actual values were very close to the predicted values line in the CRISP-DM methodology model, while the actual values were not close to the predicted values line in the OSEMN methodology, indicating that the model of CRISP-DM has in overall a better model fit. The third question is answered by the MSE values in both methodologies. Those values indicate that in one study case CRISP-DM has very high

accuracy rate regarding prediction, with an extreme small value of errors. In the second study case OSEMN methodology has higher accuracy, but the difference is not very big.

Considering the answers of second and third research questions, the results support the H1 Hypothesis demonstrating that the CRISP-DM methodology has a more accurate prediction rate than OSEMN methodology.

In the experiment conducted, we have observed that in OSEMN methodology, MSE and R-Squared values in both Iris dataset and Life Expectancy dataset remain the same. This implies a level of durability and consistency in the modeling methodology. This uniformity suggests that despite the complexity or domain-specific aspects of the datasets, the rating criteria and modeling strategies selected inside OSEMN may generalize effectively across different datasets.

While in CRISP-DM methodology, MSE and R-Squared values are different when using Iris Dataset and different when using Life Expectancy Dataset. It suggests that the dataset's properties and the modeling choices taken within CRISP-DM have an impact on the performance of the model.

This consistency might mean that OSEMN methodology is more general, as modeling techniques and measures used for evaluating their performance are consistently applied to all the datasets to assess the models ‘ability to identify more general trends. On the other hand, the variation that exists in CRISP-DM for methodology could be considered methodical and intentional and where modeling tactics are adjusted to the characteristic of a particular dataset to produce the best result in one scenario or another.

Variation in CRISP-DM metrics could suggest more adaptability in optimizing model performance across datasets, but it could also pose issues with generalizability and repeatability.

REFERENCES

[1] R. Wirth, J. Hipp, “CRISP-DM: Towards a Standard Process Model for Data Mining”, 2000

[2] Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021b). CRISP-DM Twenty years Later: From data mining processes to data science trajectories. IEEE Transactions on Knowledge and Data Engineering, 33(8), 3048–3061.https://doi.org/10.1109/tkde.2019. 2962 680

[3] Kumari, K., Bhardwaj, M., & Sharma, S. (2020b). OSEMN approach for Real time data analysis. International Journal of Engineering and Management Research, 10(02), 107–110. https://doi.org/10.31033 /ijemr.10.2.11

[4] Dineva, K., & Atanasova, T. (2018b). OSEMN PROCESS FOR WORKING OVER DATA ACQUIRED BY IOT DEVICES MOUNTED IN BEEHIVES. Current Trends in Natural Sciences. http://natsci.upit.ro/media/1641/paper-7.pdf

[5] Seventh Edition. Newtown Square: Project Management Institute.

[6] K. Nikolaidis, S. Kristiansen, Th. Plagemann, V. Goebel, K. Liestøl, M. Kankanhalli, et al., “Learning Realistic Patterns from Visually Unrealistic Stimuli: Generalization and Data Anonymization”, December 2021.

[7] Sh. Abuadbba, K. Kim, M. Kim, Ch. Thapa, S. A. Camtepe, Y. Gao, et al., “Can We Use Split Learning on 1D CNN Models for Privacy Preserving Training?”, pp. 305-318, October 2020.

[8] M. Santos, N. P. Rocha, “A Big Data Approach to Explore Medical Imaging Repositories Based on DICOM”, vol. 219, pp. 1224-1231, 2023.

[9] G. Poulis, G. Loukides, S. Skiadopoulos, A. Gkoulalas-Divanis, “Anonymizing datasets with demographics and diagnosis codes in the presence of utility constraints”, vol. 65, pp. 76-96, January 2017.

[10] J. W. T. M. deKok, M. A. Armengol de la Hoz, Y. de Jong, V. Brokke, P. W. G. Elbers, P. Thoral, et al., “A guide to sharing open healthcare data under the General Data Protection Regulation”, vol. 10, 2023.

[11] K. Abouelmehdi, A. Beni Hessane, H. Khalouf, “Big healthcare data: preserving security and privacy”, vol. 5, 2018.

[12] A. Majeed, “Attribute-centric anonymization scheme for improving user privacy and utility of publishing e-health data”, vol. 31, pp. 426-435, 2019.

[13] R. Chevrier, V. Foufi, Ch. Gaudet-Blavignac, A. Robert, Ch. Lovis, “Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review”, vol. 21, 2019.

[14] DSPA, n.d. OSEMN Data Science Life Cycle. Data Science Process Alliance. Available at: [https://www.datasciencepm.com/osemn/] (Accessed on: Oct 5, 2023).

[15] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz,. Shearer, and R. Wirth, “CRISP-DM 1.0 step-by-step data mining guide,” 2000.

[16] O. Marban, J. Segovia, E. Menasalvas, and C. Fernandez-Baizan, ´Toward data mining engineering: A software engineering approach,” Information systems, vol. 34, no. 1, pp. 87–107, 2009.

[17] IBM,“Analytics solutions unified method,”ftp://ftp. software.ibm.com /software /data/sw-library/services/ASUM.pdf, 2005.

[18] L. A. Kurgan and P. Musilek, “A survey of knowledge discovery and data mining process models,” The Knowledge Engineering Review, vol. 21, no. 1, pp. 1–24, 2006.

[19] G. Mariscal, O. Marban, and C. Fernandez, “A survey of data mining and knowledge discovery process models and methodologies,” The Knowledge Engineering Review, vol. 25, no. 02, pp. 137–166, 2010.

[20] N. Njiru and E. Opiyo, “Clustering and visualizing the status of child health in kenya: A data mining approach.” International Journal of Social Science and Technology I, 2018.

[21] N. Azadeh-Fard, F. M. Megahed, and F. Pakdil, “Variations of length of stay: a case study using control charts in the CRISPDM framework,” International Journal of Six Sigma and Competitive Advantage, vol. 11, no. 2-3, pp. 204–225, 2019.

[22] A. Daderman and S. Rosander, “Evaluating frameworks for implementing machine learning in signal processing: A comparative study of CRISP-DM, semma and kdd,” 2018.

[23] M. Rogalewicz and R. Sika, “Methodologies of knowledge discovery from data and data mining methods in mechanical engineering,” Management and Production Engineering Review, vol. 7, no. 4, pp. 97–108, 2016.

[24] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, “DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model,” Procedia CIRP, vol. 79, pp. 403–408, 2019.

[25] C. Barclay, A. Dennis, and J. Shepherd, “Application of the CRISP-DM model in predicting high school students’ examination (csec/cxc) performance,” Knowledge Discovery Process and Methods to Enhance Organizational Performance, p. 279, 2015.

[26] D. B. Fernandez and S. Luj ´ an-Mora, “Uso de la methodology CRISP-DM para guiar el proceso de miner´ıa de datos en lms,” in Tecnolog´ıa, innovacion e investigacion en los procesos de ense˜nanzaaprendizaje. Octaedro, 2016, pp. 2385–2393.

[27] L. Almahadeen, M. Akkaya, and A. Sari, “Mining student data using CRISP-DM model,” International Journal of Computer Science and Information Security, vol. 15, no. 2, p. 305, 2017.

[28] D. Oreski, I. Pihir, and M. Konecki, “CRISP-DM process model in educational setting,” Economic and Social Development: Book of Proceedings, pp. 19–28, 2017.

[29] E. Espitia, A. F. Montilla et al., “Applying CRISP-DM in a kdd process for the analysis of student attrition,” in Colombian Conference on Computing. Springer, 2018, pp. 386–401.

[30] V. Tumelaire, E. Topan, and A. Wilbik, “Development of a repair cost calculation model for daf trucks nv using the CRISPDM framework,” Ph.D. dissertation, Master’s thesis, Eindhoven University of Technology, 2015.

[31] F. Schafer, C. Zeiselmair, J. Becker, and H. Otten, “Synthesizing ¨ CRISP-DM and quality management: A data mining approach for production processes,” in 2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD). IEEE, 2018, pp. 190–195.

[32] E. G. Nabati and K.-D. Thoben, “On applicability of big data analytics in the closed-loop product lifecycle: Integration of CRISPDM standard,” in IFIP International Conference on Product Lifecycle Management. Springer, 2016, pp. 457–467.

[33] H. Nagashima and Y. Kato, “Aprep-dm: a framework for automating the pre-processing of a sensor data analysis based on CRISPDM,” in 2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops). IEEE, 2019, pp. 555–560.

[34] S. B. Gomez, M. C. G ´ omez, and J. B. Quintero, “Inteligencia de ´ negocios aplicada al ecoturismo en colombia: Un caso de estudioInformation Systems and Technologies, CISTI 2019. IEEE Computer Society, 2019, p. 8760802.

[35] R. Ganger, J. Coles, J. Ekstrum, T. Hanratty, E. Heilman, J. Boslaugh, and Z. Kendrick, “Application of data science within the army intelligence warfighting function: problem summary and key findings,” in Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, vol. 11006. International Society for Optics and Photonics, 2019, p. 110060N.

[36] R. P. Bunker and F. Thabtah, “A machine learning framework for sport result prediction,” Applied computing and informatics, 2017.

[37] R. Barros, A. Peres, F. Lorenzi, L. K. Wives, and E. H. da Silva Jaccottet, “Case law analysis with machine learning in Brazilian court,” in International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer, 2018, pp. 857–868.

[38] K. J. Cios, A. Teresinska, S. Konieczna, J. Potocka, and S. Sharma, “A knowledge discovery approach to diagnosing myocardial perfusion,” Engineering in Medicine and Biology Magazine, IEEE, vol. 19, no. 4, pp. 17–25, 2000.

[39] K. J. Cios and L. A. Kurgan, “Trends in data mining and knowledge discovery,” in Advanced techniques in knowledge discovery and data mining. Springer, 2005, pp. 1–26.

[40] S. Moyle and A. Jorge, “Ramsys-a methodology for supporting rapid remote collaborative data mining projects,” in ECML/PKDD 2001 Workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning: Internal SolEuNet Session, 2001, pp. 20–31.