**THE SUPERIOR COLLEGE LAHORE**

|  |
| --- |
| chk_captcha |

**Faculty of Computer Science & IT**

**Department of Software Engineering**

**Final Year Project**

**PROJECT REPORT (Part-1)**

**Detecting Influential Nodes Using Multilayered Social Network Analysis (SNA)**

Project ID: **[write ID here Issued by FYP Manager]**

**Project Team**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Student Name** | **Student ID** | **Program** | **Contact Number** | **Email Address** |
| Zainab Akhter | Bcsm-s17-047 | BSCS | 03058650804 | Zainab.akhter661@gmail.com |
| Husnain Ashraf | Bcsm-s17-061 | BSCS | 03486618536 | ali.husnain5827@gmail.com |
| Fahad Rajput | Bcsm-s17-021 | BSCS | 03164254105 | Fahadrajput115@gmail.com |
|  |  |  |  |  |
|  |  |  |  |  |

**Project Supervisor: Jr. Lec Muhammad Kaleem Ullah**

**Designation: Jr. Lec**

**Project Report**

**Detecting Influential Nodes Using Multilayered Social Network Analysis (SNA)**

**Change Record**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Author(s)** | **Version** | **Date** | **Notes** | **Supervisor’s Signature** |
|  | 1.0 |  | <Original Draft> |  |
|  |  |  | <Changes Based on Feedback from Supervisor> |  |
|  |  |  | <Changes Based on Feedback From Faculty> |  |
|  |  |  | <Added Project Plan> |  |
|  |  |  | <Changes Based on Feedback from Supervisor> |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**APPROVAL**

|  |  |
| --- | --- |
| **Project Supervisor** | |
| Comments: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | | |
| Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  | |
| Date:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | Signature:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | |

|  |  |
| --- | --- |
| **Project Manager** | |
| Comments: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | |
|  |  |
| Date:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | Signature:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

|  |  |
| --- | --- |
| **Head of the Department** | |
| Comments: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | |
|  |  |
| Date:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | Signature:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

# Dedication

*I dedicate this work to Allah Almighty my creator, my strong pillar, my source of inspiration, wisdom, knowledge and understanding. He has been the source of my strength throughout this program and on His wings only have I soared. I also dedicate this work to my Supervisor (Muhammad Kaleem Ullah) who has encouraged me all the way and whose encouragement has made sure that I can finish that I have started. He always supported me in my educational career. I also dedicate this work to my group mates Husnain Ashraf & Fahad Rajput who helped me a lot. Thanks all of you.*

# Acknowledgements

Firstly, My deepest gratitude goes to Allah who has provided all that was needed to complete this project and the program for which it was undertaken. There was never lack or want. Throughout this entire study, He took care of everything that would have stopped me in my tracks and strengthened me even through my most difficult times.

Secondly, I would also like to express my special thanks of gratitude to my supervisor (Muhammad Kaleem Ullah) as well who gave me the excellent opportunity to do this wonderful project. Which also helped me in doing a lot of Research and i came to know about so many new things I am thankful to him.

Lastly, I would also like to thank my parents and group mates Husnain Ashraf and Fahad Rajput who helped me a lot in finishing this project within the limited time. I am making this project not only for marks but to also increase my knowledge. Thanks again to all who helped me.

# Executive Summary

Along with the explosive growth of the phenomenon Online Social Networks (OSN), identifying

influential users in OSNs’ has received a great deal of attention. However, the development of

practical approaches for identifying them is still in its infancy. By means of a structured literature review, our research will analyze and synthesize the publications particularly from two perspectives. From a research perspective, they find that existing approaches mostly build on users connectivity and activity but hardly consider further characteristics of influential users. Moreover, we outline two major research streams. It becomes apparent that most marketing oriented articles draw on real-world data of OSN, while more technology-oriented papers rather have a theoretical approach and mostly evaluate their artifacts by means of formal proofs. This research will find that a stronger collaboration between the scientific Business and Information Systems Engineering (BISE) and Marketing communities could be mutually beneficial. With respect to a practitioner’s perspective, they compile advice on the practical application of approaches for the identification of influential users. The validity of proposed system will be tested with the result of previous analyzed social networks. So we have made a methodology which will use SNA methodologies to enhance business by detecting the key Players, developing the marketing campaigns, detecting behaviors of customers, predicting the customer’s churns, developing recommender systems etc.

# Table of Contents

[Dedication iv](#_Toc505638416)

[Acknowledgements v](#_Toc505638417)

[Executive Summary vi](#_Toc505638418)

[Table of Contents vii](#_Toc505638419)

[List of Figures ix](#_Toc505638420)

[List of Tables x](#_Toc505638421)

[Chapter 1 1](#_Toc505638422)

[Introduction 1](#_Toc505638423)

[1.1. Background 2](#_Toc505638424)

[1.2. Motivations and Challenges 2](#_Toc505638425)

[1.3. Goals and Objectives 2](#_Toc505638426)

[1.4. Literature Review/Existing Solutions 2](#_Toc505638427)

[1.5. Gap Analysis 2](#_Toc505638428)

[1.6. Proposed Solution 2](#_Toc505638429)

[1.7. Project Plan 3](#_Toc505638430)

[1.7.1. Work Breakdown Structure 3](#_Toc505638431)

[1.7.2. Roles & Responsibility Matrix 3](#_Toc505638432)

[1.7.3. Gantt Chart 3](#_Toc505638433)

[1.8. Report Outline 3](#_Toc505638434)

[Chapter 2 4](#_Toc505638435)

[Software Requirement Specifications 4](#_Toc505638436)

[2.1. Introduction 5](#_Toc505638438)

[2.1.1. Purpose 5](#_Toc505638439)

[2.1.2. Document Conventions 5](#_Toc505638440)

[2.1.3. Intended Audience and Reading Suggestions 5](#_Toc505638441)

[2.1.4. Product Scope 5](#_Toc505638442)

[2.1.5. References 6](#_Toc505638443)

[2.2. Overall Description 6](#_Toc505638444)

[2.2.1. Product Perspective 6](#_Toc505638445)

[2.2.2. Product Functions 6](#_Toc505638446)

[2.2.3. User Classes and Characteristics 6](#_Toc505638447)

[2.2.4. Operating Environment 7](#_Toc505638448)

[2.2.5. Design and Implementation Constraints 7](#_Toc505638449)

[2.2.6. User Documentation 7](#_Toc505638450)

[2.2.7. Assumptions and Dependencies 7](#_Toc505638451)

[2.3. External Interface Requirements 8](#_Toc505638452)

[2.3.1. User Interfaces 8](#_Toc505638453)

[2.3.2. Hardware Interfaces 8](#_Toc505638454)

[2.3.3. Software Interfaces 8](#_Toc505638455)

[2.3.4. Communications Interfaces 9](#_Toc505638456)

[2.4. System Features 9](#_Toc505638457)

[2.4.1. System Feature 1 9](#_Toc505638458)

[2.4.1.1. Description and Priority 9](#_Toc505638459)

[2.4.1.2. Stimulus/Response Sequences 9](#_Toc505638460)

[2.4.1.3. Functional Requirements 9](#_Toc505638461)

[2.4.2. System Feature 2 10](#_Toc505638462)

[2.4.2.1. Description and Priority 10](#_Toc505638463)

[2.4.2.2. Stimulus/Response Sequences 10](#_Toc505638464)

[2.4.2.3. Functional Requirements 10](#_Toc505638465)

[2.4.3. System Feature 3 (and so on) 11](#_Toc505638466)

[2.5. Other Nonfunctional Requirements 11](#_Toc505638467)

[2.5.1. Performance Requirements 11](#_Toc505638468)

[2.5.2. Safety Requirements 11](#_Toc505638469)

[2.5.3. Security Requirements 12](#_Toc505638470)

[2.5.4. Software Quality Attributes 12](#_Toc505638471)

[2.5.5. Business Rules 12](#_Toc505638472)

[2.6. Other Requirements 12](#_Toc505638473)

[Chapter 3 13](#_Toc505638474)

[Use Case Analysis 13](#_Toc505638475)

[3.1. Use Case Model 14](#_Toc505638476)

[3.2. Fully Dressed Use Cases 14](#_Toc505638477)

[Chapter 4 15](#_Toc505638478)

[System Design 15](#_Toc505638479)

[4.1. Architecture Diagram 16](#_Toc505638480)

[4.2. Domain Model 16](#_Toc505638481)

[4.3. Entity Relationship Diagram with data dictionary 16](#_Toc505638482)

[4.4. Class Diagram 17](#_Toc505638483)

[4.5. Sequence / Collaboration Diagram 17](#_Toc505638484)

[4.6. Operation contracts 17](#_Toc505638485)

[4.7. Activity Diagram 18](#_Toc505638486)

[4.8. State Transition Diagram 18](#_Toc505638487)

[4.9. Component Diagram 18](#_Toc505638488)

[4.10. Deployment Diagram 19](#_Toc505638489)

[4.11. Data Flow diagram [only if structured approach is used - Level 0 and 1] 19](#_Toc505638490)

[Chapter 5 20](#_Toc505638491)

[Implementation 20](#_Toc505638492)

[5.1. Important Flow Control/Pseudo codes 21](#_Toc505638493)

[5.2. Components, Libraries, Web Services and stubs 21](#_Toc505638494)

[5.3. Deployment Environment 21](#_Toc505638495)

[5.4. Tools and Techniques 22](#_Toc505638496)

[5.5. Best Practices / Coding Standards 22](#_Toc505638497)

[5.6. Version Control 22](#_Toc505638498)

[Appendices 23](#_Toc505638499)

[Appendix A: Information / Promotional Material 24](#_Toc505638500)

[Reference and Bibliography 27](#_Toc505638502)

[Index 29](#_Toc505638503)

# List of Figures

1.1 Caption of first figure of first chapter 6

1.2 Caption of second figure of first chapter 7

2.1 Caption of first figure of second chapter 14

2.2 Caption of second figure of second chapter 22

2.3 Caption of third figure of second chapter 26

5.1 Caption of first figure of fifth chapter 49

5.2 Caption of second figure of fifth chapter 49

# List of Tables

1.1 label of first table of first chapter 6

1.2 label of second table of first chapter 7

2.1 label of first table of second chapter 14

2.2 label of second table of second chapter 22

2.3 label of third table of second chapter 26

5.1 label of first table of fifth chapter 49

5.2 label of second table of fifth chapter 49

# Chapter 1

# Introduction

**Chapter 1:** Introduction

This chapter is all about basics and pre inputs of our project. In this chapter we will discuss about our project’s important aspects these are background of the project, motivations, challenges that we will face through out the processing state, our main goal and objectives, we will discuss deeply about the existing solutions, Gap Analysis that tells us to get from our current state to our desired state then we will discuss the proposed solution our methodology that will tell us what are benefits of our methodology who will get the real benefit through it solution to the problem and way to implement it, through the Project plan, WBS, RACI and Gantt we will dicuss the Performance, Responsibilities and Roles of the project team. In a nutshell we will give the Project Report.

## Background

Like many other fields, Online Business Community have attracted huge amount of research,

especially from Information Technologies and networks due to various reasons. First of all Online Businesses have climbed up the graph since the recent years because of the increasing number of digital buyers across the globe. The digital buyers had increased from 1.32 Billion to 1.92 Billion in just few years i.e. 2014 to 2019 respectively. According to researchers the amount is expected to rise to $2.14 Billion at the end of 2021. Companies nowadays are using the social network analysis tools and techniques to enhance their business positions. Analyzing the social networks provide the massive amount of useful information for the business. The applications of Social Network Analysis with respect to Business include Recommender System, Churn Prediction, Target Marketing and many more. Because of this, there is a room for effective research in order to contribute in the area of online business communities. In the recent years, the business communities have started using the SNA methodologies to enhance their business by detecting the key Players, developing the marketing campaigns, detecting behaviors of their customers, predicting the customer’s churns, developing recommender systems etc.

## Motivations and Challenges

Thing which motivates us is that fact, people who are hardworking and doing amazing work but they don’t know how to spread their little business and they don’t know how to make more money by their business. The major Challenge here is to make people able to use our methodology, some people do not have much exposure to smartphones/Laptops so it is a challenge for us to make people use our methodology and make them believe that it is for their benefit. We have to market our methodology wisely to get most out of it. By using our this our users can enhance their business by detecting the key Players, developing the marketing campaigns, detecting behaviors of their customers, predicting the customer’s churns, developing recommender systems etc.

## Goals and Objectives

The aim of this project is to perform social network analysis on multi layered networks to detect Influential nodes which can help marketing community in improving their businesses. Our ambition is to connect our society’s skill full people to the bigger market and outer world, we have so many people in our country who have skills but they cannot get best out of it. Our goal is to provide a better life to those who actually do work but did not get most of the credit. Those people need to be in limelight. They cannot afford a fancy shop or marketing plan to sell their products on a broader scale so we are focusing on that part of our society to strengthen them.

**Objectives**

* To gather social network datasets
* To design a generic feature vector for cross network analysis
* To develop classification techniques for analysis (decision tree, random forest, id3, naive bayes Algorithms)
* To extract influential nodes using social network analysis

## Literature Review/Existing Solutions

A number of researchers had worked on the social network analysis for different kind of the

applications. Kimura Metal presents a cascaded method for detection of most key players

in online social networks. Authors suggested different algorithms and techniques that extract

network parameters which play a very vital role in detecting the key players in social network.

Based on their statistical analysis we can say, their proposed algorithm is highly recommended

than the greedy solution of the Leicht M et al and Newman community structure algorithm.

Salvatore Catanese et al describes the friendship relation between users and how to access

data of Facebook users. For achieving the purpose they used the web crawlers to access data

directly from the website. By using gathered data and information the author constructed a sub

group graph that represents the anonymous relationship among a sub group of significant users. An adhoc privacy complaint crawlers is studied to extract data from Facebook. Rejection sampling methods and Breadth-first-search (BFS) are used to minimize the biasness and for

visualizing the structural characteristics of different samples that consist of huge no. of key

players. Authors developed a visualizing tool for analysis of qualitative and quantitative

characteristics of social media. To achieve the results they improved the efficiency of existing

online social media analysis (OSN) techniques and adopted existing techniques methods and

algorithms. Pasquale De Moeetal describes that how to analyze the behaviors of new users

to predict whether the two nodes could be considered as a similar one. Author proposed a

framework where the estimation can be performed to verify and check the similarities of two

users that is based on visualizing the different activities that includes social events in which users are fully involved and on the basis of information of social relation i.e. common groups of users and friends. Leidys del Carmen measured the clustering and association rules with famous CRISP. DM method to analyze the behaviors of the customers of the fashion industry in Instagram social network, which provided the industry with the handful of important information regarding their products and their trends and likeness among their customers.Numerous related research discussed the presence of influential nodes in a particular social networks (e.g. Facebook, Twitter, Micro blog). Cha et al extracted Twitter data and did analysis to find the influence of Twitter users by comparing the network metrics values of Clustering Coefficient, Degree, Eigenvector Centrality and retweets. Rossi, Vazirgiannis and Malliaros proposed a framework that visualize the complex social networks and detect influential nodes in a network. Authors proposed a famous technique i.e. K-truss composition method that helps in visualizing and analyzing the social networks and detection of Influential nodes.

## Gap Analysis

In today’s modern world, everyone is competing to achieve their target, but there is a gap between their target and achievement. A Gap Analysis helps a project team to understand what they want to be, the gap between where they are now and where they want to be and hence, what steps should be taken to close the gap.

**Current State**

In our project we are at the initial state of the performance we are doing deeply research about the whole project this is our current state where we are now.

**Gap Analysis( How do we do that?/ What do we need to do to get there? )**

Our Proposed schema is to detect the Key Players with the graphic matrix by using python language then apply the classification methods for the evaluation and detection of the key Player our schema is divided into 4 phases.

1. Social Network Selection (Dataset)

2. Pre-Processing of the Data

3. Visualization and Feature Selection

4. Classification

**Desired State (Where we want to go?)**

Finally we will reach to the final state that is our Desired State. The aim of this project is to perform social network analysis on multi layered networks to detect Influential nodes which can help marketing community in improving their businesses.

## Proposed Solution

Our Proposed schema is to detect the Key Players with the graphic matrix by using python language then apply the classification methods for the evaluation and

detection of the key Player our schema is divided into 4 phases.

1. Social Network Selection (Dataset)

2. Pre-Processing of the Data

3. Visualization and Feature Selection

4. Classification

## Project Plan

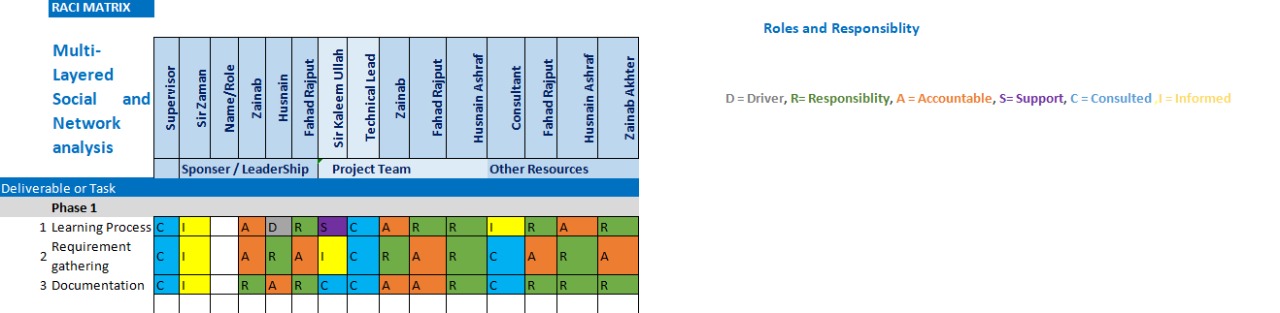
Data Gathering is the first phase then our next phase is Data Preprocessing (Selection of AI algorithms) then we will perform Training and Testing.

## Work Breakdown Structure

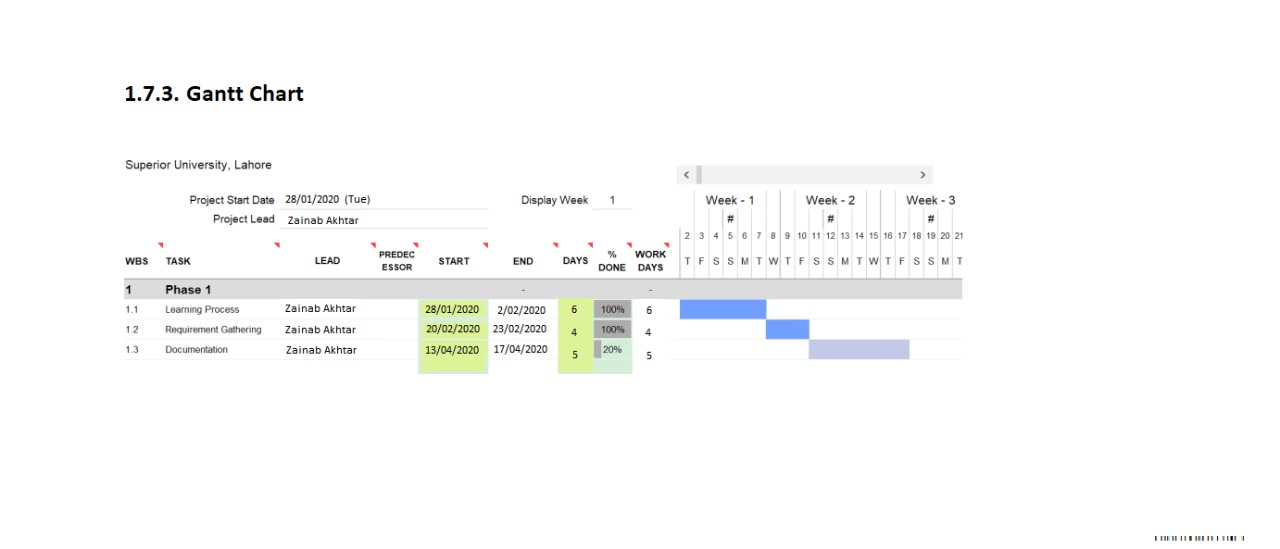
We have Break down all the processes involved in the development of our project to understand the project in a better way and to develop the project on these guidelines, after all the processes we will combine them to form our functional result.



## Roles & Responsibility Matrix



## Gantt Chart



## Report Outline

Documentation is currently in progress.

# Chapter 2

# Software Requirement Specifications

**Chapter 2:** Software Requirement Specifications



## Introduction

## Purpose

The purose of this project is to perform social network analysis on multi layered networks to detect Influential nodes which can help marketing community in improving their businesses. By using this our users can enhance their business by detecting the key Players, developing the marketing campaigns, detecting behaviors of their customers, predicting the customer’s churns, developing recommender systems etc. As we know, the human brain grasps visuals more easily than table reports. Data visualizations allow decision makers to be notified quickly of new data insights and take necessary actions for business growth.

## 

## Document Conventions

This document features some terminology which readers may be unfamiliar with. This document follows MLA format. Bold-faced text has been used to emphasize section and sub section headings. Highlighting is to point out words in the glossary and italicized text is used to label and recognize diagrams.

## 

## Intended Audience and Reading Suggestions

This document is to be read by the Development Team, the Project Managers, Marketing Staff and Testers.

**Developers** who can review project’s capabilities and more easily understand where their efforts should be targeted to improve or add more features to it.

**Project testers** can use this document as a base for their testing strategy as some bugs are easier to find using a requirements document. This way testing becomes more methodically organized.

**End users** of this methodology who wish to read about what this project can do.

**Introduction**: This section offers a summary of the project, including goals and objectives, project scope, general system details, and some major constraints associated with the intended platform.

**Overall Description:** Marketing staff have to become familiar to the various product features in order to effectively advertise the product.

**System features:** Testers need an understanding of the system features to develop Meaningful test cases and give useful feedback to the developers.

**System Design:** Readers interested in how this project organizes and handles data should consult this section, which covers data structures and flow patterns utilized by the system.

**Architectural Diagram:** This section describes our methodology system class by class, including interface details, class hierarchies, performance/design constraints, process details, and algorithmic models.

**Appendices:** This section includes any additional information which may be helpful to readers.

## Product Scope

We are doing social network analysis on multi layered networks to detect Influential nodes by visualization, higher centrality and classifiers, which can help marketing community in improving their businesses. By using this our users can enhance their business by detecting the key Players, developing the marketing campaigns, detecting behaviors of their customers, predicting the customer’s churns, developing recommender systems etc. Detection of the key player in a social network is an old and well-known problem, it is no different than finding a needle in a haystack. Social network analysis is being used in the modern era for various applications some of them include; fraud detection, terrorist activities in covert networks, military surveillance for enemy activities. Humans more easily grasp information through visualization. Data visualization helps business stakeholders analyze reports regarding sales, marketing strategies, and product interest. Based on the analysis, they can focus on the areas that require attention to increase profits, which in turn makes the business more productive.

* In a business context, visualization helps convey a story to decision makers, allowing them to act more quickly than if the data were presented as reports.
* Handling large amount of data in a pictorial format to provide a summary of unseen patterns in the data, revealing insights and the story behind the data to establish a business goal.
* Visualizing business data to manage growth and converting trends into business strategies by making sense of your information.
* Revealing previously unnoticed key points about the data sources to help decision makers compose data analysis reports.

## References

* Kimura, M., Yamakawa, K. ; Saito, K. ; Motoda, H.," CommunityAnalysis of Influential Nodes for Infonnation Diffusion on a SocialNetwork" IEEE International Joint Conference on Neural Networks, pp.1358 - 1363, 2008.
* Salvatore Catanese et al., ―Extraction and Analysis of FacebookFriendship Relations,‖ Springer, 291-324, 2012.
* Pasquale De Meo et al., ―Finding Similar Users in Facebook,‖Information Science Reference, 304-323, 2012.
* Leidys del Carmen, Analysis of the Behavior of Customers in the Social Networks Using Data Mining Techniques, Doi: 10.1109/ASONAM.2016.7752301
* LR CHA M Y, HADDADI H, BENEVENUTO F, et al. Measuring userinfluence in twitter: The million follower fallacy. Proceedings of the4th International AAAI Conference on Weblogs and Social Media,2010: 10-17.
* Maria-Evgenia G. Rossi, Fragkiskos D. Malliaros, MichallsVazirginnis. Spread it Good, Spread it Fast: Identification of InfluentialNodes in Social Networks. Proceedings of the 24th InternationalConference on World Wide Web, pages 101-102.

## Overall Description

## Product Perspective

Our timesaving for those market communities who like to enhance their business by detecting the key Players, developing the marketing campaigns, detecting behaviors of their customers, predicting the customer’s churns, developing recommender systems etc. It will help business stakeholders analyze reports regarding sales, marketing strategies, and product interest. Based on the analysis, they can focus on the areas that require attention to increase profits, which in turn makes the business more productive.methodology is a very useful and

## Product Functions

Our methodology provides users the following functions:

* Social Network Selection (Dataset)
* Pre-Processing of the Data
* Design a generic feature vector for cross network analysis
* Visualization and Feature Selection
* Develop classification techniques for analysis
* Extract influential nodes using social network analysis

## User Classes and Characteristics

* Marketing community for improving their businesses by detecting influential nodes.
* Decision makers, understand how the business data is being interpreted to determine business decisions.
* Typical Users, such as students, who want to use this methodology for analyzing networks (Social networks, Social Media networks, etc.)
* Advanced/Professional Users, such as engineers or researchers, who want to use our methodology for more demanding graph analysis.
* Programmers who are interested in working on the project by further developing it or fix existing bugs.

## Operating Environment

* Windows
* Mac OS X
* PyCharm

## Design and Implementation Constraints

Our methodology is developed in Python and therefore requires Python to be installed on the user’s system, it uses Numpy, Pandas, Matplotlib and Scikit Libraries. It will use the centrality and it’s types then it will use Artificial Intelligence Algorithms like decision tree, random forest, id3, naive bayes to find classification techniques for analysis.

## Assumptions and Dependencies

Our methodology is developed in Python and therefore requires Python to be installed on the user’s system. It requires Python latest version 3.8.3 for better performance. This applies to Windows and Mac OS X users.

## External Interface Requirements

## Software Interfaces

#### 1. PyCharm

It is one of the best and extensively used IDE, which is created by Jet Brains. This IDE is used by developers for developing the productive Python and develops a neat & maintainable code. It supports:

* Integrated tools for unit testing, code inspection, version control, refactoring, and navigation.
* Support for libraries like scikit-learn, matplotlib, NumPy, and pandas.
* Ability to plot, manage, and explore graphs in real time.
* Support for database languages like SQL (via plugins).

#### 2. Spyder

It is another open-source IDE that can be [used as a python compiler](https://www.educba.com/python-compilers/) for python code development. It supports:

* Support for libraries like NumPy, SciPy, Matplotlib, and others.
* Interactive console for building and testing applications.
* Simple and lightweight.
* Multi-language support.
* Auto-completion, syntax highlighting, and resource management.

#### 3. Rodeo

It is one of the [best python IDE’s](https://www.educba.com/python-ide-on-linux/) built for work related to data science such as extracting information from various tools and mapping for issues. It supports:

* Auto-completion.
* Syntax highlighting.
* Support for Vim and Emacs text editors (cross platform).
* Integrated tutorials for Python beginners.
* Environment pane for keeping track of variables and functions.
* Data pane for managing files, photos, packages, and settings.

### 4. Jupyter Notebook

[Jupyter Notebook](https://academy.vertabelo.com/blog/jupyter-notebook-python-ide-installation-tips/) is popular open-source Python IDE among data scientists. It supports:

* Data cleaning
* Data transformation
* Numerical simulation
* Statistical modeling
* Data visualization
* Machine learning

**2.3.4. Communications Interfaces**

Our methodology requires an internet connection to update some of its components and modules.

## System Features

This section demonstrates most prominent functions and explains how they can be used and the results they will give back to the user.

## Data preprocessing

 The purpose of this step is to clean the data in order to facilitate further steps.

## Description and Priority

Preprocessing step checks for duplicate entries and removes all such entries to avoid redundancy and to handle missing values in the data. The preprocessing technique identifies the missing feature values and then they are replaced by the mean value for that feature. This procedure is performed for those attributes where values are missing in less than 50% of the instances. If the number of instances with missing values is more than or equal to 50%, the particular attribute is rejected and not used further. It has higher priority because if data will not filter then with next steps we will not able to achieve accurate results.

## Stimulus/Response Sequences

First step uses two rank sum tests, that is, Wilcoxon rank-sum and Ansari-Bradley tests. Wilcoxon rank-sum test is a nonparametric test of the null hypothesis that two populations are the same against an alternative hypothesis that the two distributions differ only with respect to the median. Ansari-Bradley test compares two independent samples which come from the same distribution against the alternative that they come from the same distributions having the same median and shape but different variances. Preprocessing step also checks for duplicate entries and removes all such entries to avoid redundancy. The last step in preprocessing is to handle missing values in the data. The preprocessing technique identifies the missing feature values and then they are replaced by the mean value for that feature. This procedure is performed for those attributes where values are missing in less than 50% of the instances. If the number of instances with missing values is more than or equal to 50%, the particular attribute is rejected and not used further.

## Functional Requirements

REQ-SF1-1: **Redundant Feature Removal**, It is very important to remove redundant features from the given data.

REQ-SF1-2: **Removal of duplicate entries**, It is very important to check the duplicate entries and removes all such entries to avoid redundancy.

REQ-SF1-3: **Handling missing values**, It is very important to handle missing values in the data. It identifies the missing feature values and then they are replaced by the mean value for that feature.

## Key Player Detection

This step will find the key players from the given dataset on multi layered networks.

## Description and Priority

Detection of the key player in a social network is an old and well-known problem, it is no different than finding a needle in a haystack. Key players normally appear as most central nodes in any network. It has a very high priority to analyze the customer behaviors and relative users for marketing campaigns and improving their business.

## Stimulus/Response Sequences

The proposed framework for key player detection consists of centrality measures. Key players normally appear as most central nodes in any network; so, they have significant values of centrality measures. The four centrality measures which we have included in our proposed model are degree centrality (DC), betweenness centrality (BC), closeness centrality (CC), and eigenvector centrality (EC).

## Functional Requirements

REQ-SF2-1: **Degree Centality** is a simple count of the total number of connections linked to a vertex. It can be thought of as a kind of popularity measure.

REQ-SF2-2: **Betweeness Centrality** captures how much a given node is in-between others. This metric is measured with the number of shortest paths between any couple of nodes in the graphs that passes through the target node.

R*EQ-SF2-3:* **Closeness Centrality**is a way of detecting nodes that are able to spread information very efficiently through a graph. The closeness centrality of a node measures its average farness (inverse distance) to all other nodes. Nodes with a high closeness score have the shortest distances to all other nodes.

REQ-SF2-3: **Eigenvector Cenrality** is a more sophisticated view of centrality. It measures a node’s importance while giving consideration to the importance of its neighbors

## Classifiers

The proposed framework for key player detection consists of centrality measures for each node followed by classifiers for accurate detection of key players. These are the algorithms for training and testing data. Classification belongs to the category of supervised learning.

## Description and Priority

Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function from input variables to discrete output variables. Classification belongs to the category of supervised learning where the targets also provided with the input data. It has a very high priority in our proposed schema because of classifiers we can detect more accurate key layers and train our data. We can perform training and testing through these classifiers

* + - 1. **Stimulus/Response Sequences**

Once all nodes are represented by feature vectors, next phase is to classify them as key player or normal member. We will use Decision tree, Random forest, id3 and Naive bayes classifers. The purpose of using these four classifiers is to accurately model the distribution of data and to find accurate decision boundary by using the strengths of all four classifiers.

## 2.4.3.3 Functional Requirements

REQ-SF3-1: **Decision Tree** builds classification or regression models in the form of a tree structure. It utilizes an if-then rule set which is mutually exclusive and exhaustive for classification. The rules are learned sequentially using the training data one at a time. Each time a rule is learned, the tuples covered by the rules are removed. This process is continued on the training set until meeting a termination condition.

REQ-SF3-2: **Random Forest** like its name implies, consists of a large number of individual decision trees that operate as an [ensemble](https://en.wikipedia.org/wiki/Ensemble_learning). Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction

REQ-SF3-3: **id3 i**n Decision Tree learning, one of the most popular algorithms is the **ID3 algorithm** or the **Iterative Dichotomiser 3 algorithm. We will use it** to generate a Decision Tree from a dataset

REQ-SF4-4: **Naive Bayes** classifier is a probabilistic machine learning model we will use it for classification task.

## Other Nonfunctional Requirements

## Performance Requirements

Our methodology requires a system with a fast CPU speed, 4 GB of RAM and a best compatible graphics card. However, these requirements can support effectively graphs of up to 1000 nodes and edges. Performance depends on the graph size and as a result, the system requirements for bigger graphs are more demanding. Few more performance requirements are:

* Display the relevant facts.
* Hide unwanted details in your visuals.
* Use less data by using filters.
* Provide pleasant colors to the graphs for better understanding.
* Categorize large data with drill-down behavior.
* Auto-refresh your visuals to show the most recent data.
* Don’t show too much information in your visuals.
* Response Time of the system after user interaction must be minimum.
* System must be ready to tackle large amount of dataset at single time.
* In one respect, scalability is simply specified as increase in the system’s workload that the system should be able to process.

## Safety Requirements

To ensure that no one of our users loses any data while using our methodology due to a crash or a bug of some kind, the developer team updates this methodology regularly.

## Security Requirements

Our work does not have any security requirements and thus any type of user can use it without any additional privileges.

## Software Quality Attributes

We provide the users with both simple and advanced functions. These are easy to use, it can be used by both experts and typical users. However, users must already have a basic knowledge of graphs before using it.

**Usability:** Our methodology must be easy and usable for all type of users.

**Correctness**: Data and statistics provided by our work must be correct.

**Availability**: Our methodology must be available in every situation even if there is a heavy workload.

**Reliability**: Our system fulfills a function for a specified number of input trials under specified input conditions in a specified time interval.

## Business Rules

In case of any failure the developers will respond quickly and try to capture the error.

## Other Requirements

* Display the relevant facts.
* Hide unwanted details in your visuals.
* Use less data by using filters.
* Provide pleasant colors to the graphs for better understanding.
* Categorize large data with drill-down behavior.
* Auto-refresh your visuals to show the most recent data.
* Don’t show too much information in your visuals.

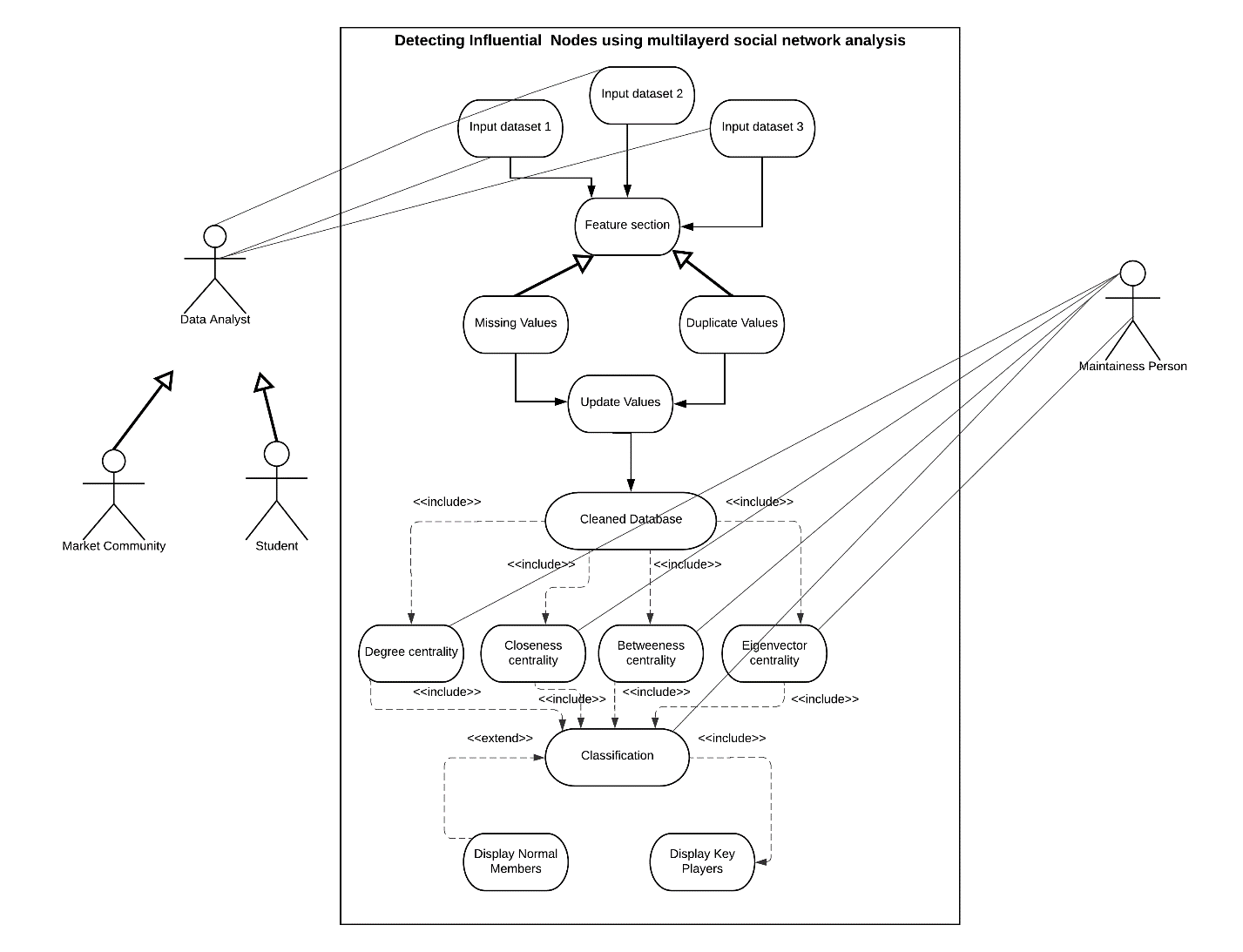
# Chapter 3

# Use Case Analysis

**Chapter 3:** System Analysis

This chapter is all about the detailed of our project. It is the core for End users of this methodology who wish to read about what this project can do. In this chapter we will discuss our project in detail. We will see use case diagram that will show the whole procedure of our methodology and then we will see the fully dressed use case by this we will able to find out in detail that how our project will work. This part is the core part for understanding the functionality of project.

## Use Case Model



**Maintenance Person: Correct spellings of second actor.**

## Fully Dressed Use Cases

|  |  |
| --- | --- |
| **Use Case title** | Detecting Influential Nodes using multilayerd social network analysis. |
| **Description** | This use case allows user to enter datasets and find key players from the dataset through preprocessing, Higher centrality and classification. |
| **Primary Actor** | Data Analyst |
| **Secondary Actor** | Maintenance Person |
| **Stackholders** | * Data Analyst * Maintenance Person * Testers * Developers |
| **Preconditions** | * User must have a laptop/PC. * User must have to input datasets. |
| **Postconditions** | The system displays the relevant results. |
| **Main Success scenario (basic flow)** | * User input datasets * User must analyse visuals * The system displays data in graph and tables. * The system displays the key players. * The use case ends. |
| **Extensions** | * User inputs corrupt datasets. * If the system doesn’t show the accurate visuals. |
| **Special Requirements** | * Display the relevant facts. * Use less data by using filters. * **Usability:** Our methodology must be easy and usable for all type of users. * **Correctness**: Data and statistics provided by our work must be correct. * **Availability**: Our methodology must be available in every situation even if there is a heavy workload. * **Reliability**: Our system fulfills a function for a specified number of input trials under specified input conditions in a specified time interval. |

# Chapter 4

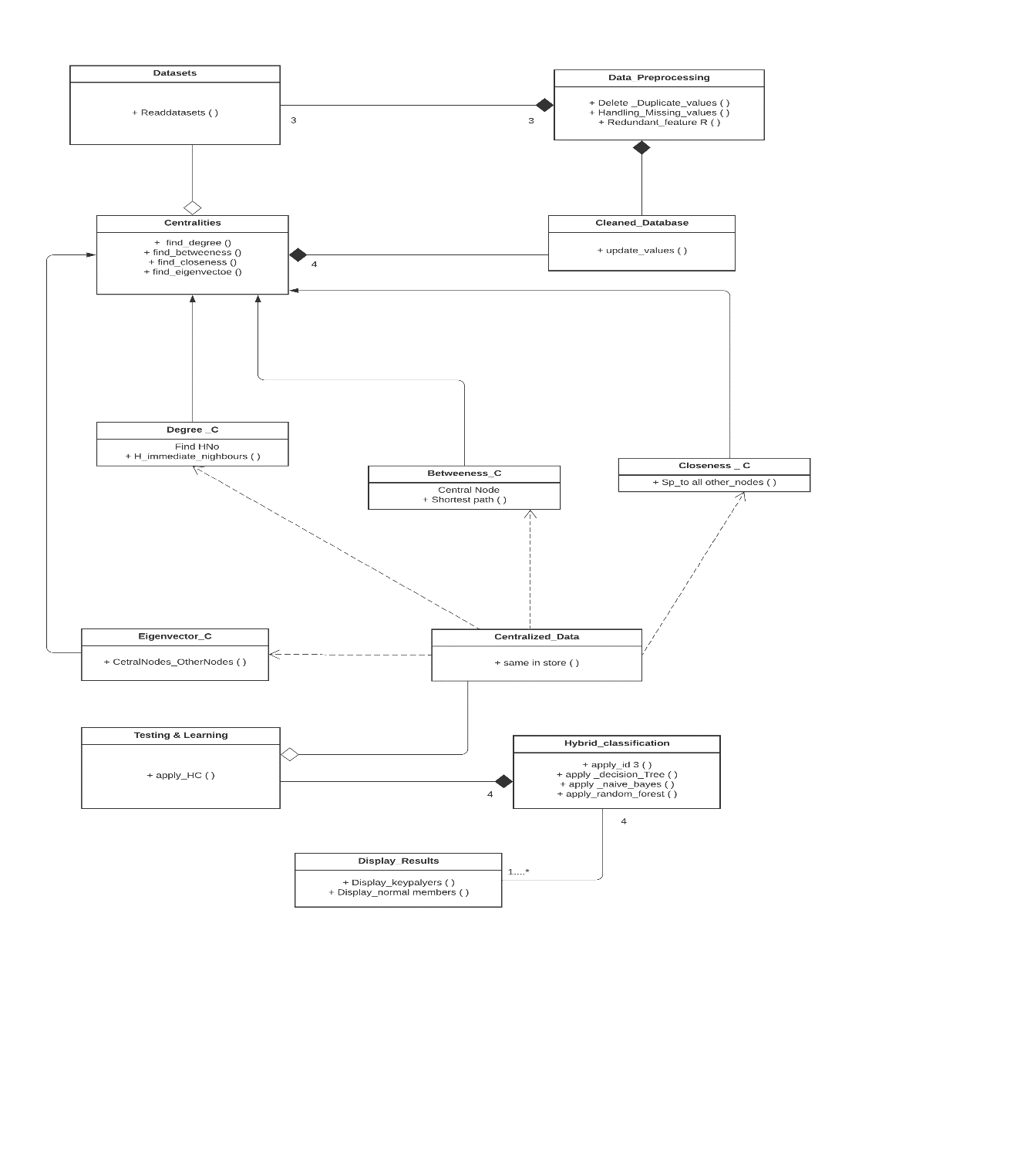
# System Design

**Chapter 4:** System Design

System design is the process of defining the architecture, modules, interfaces, and data for a system to satisfy specified requirements. This chapter will be all about system design of our methodology. What would be the architecture and what modules we are dealing with. In this chapter we will see the diagrams and these diagrams will tell the whole procedure efficiently. This phase will focus on the solution domain that is “**How to implement?*”***

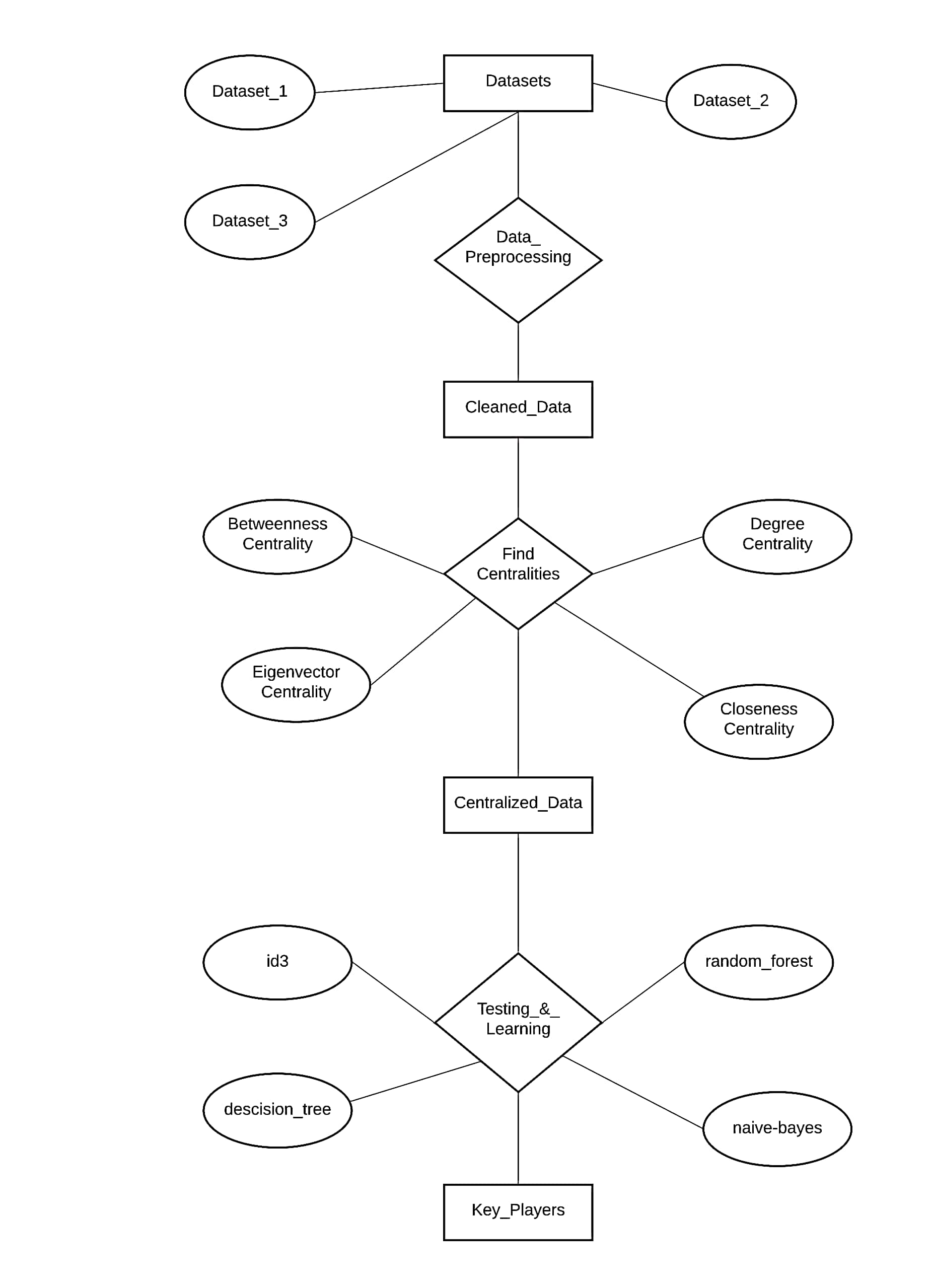
## Domain Model

A domain model is a system of abstractions that describes selected aspects of a sphere of knowledge, influence or activity the model can then be used to solve problems related to that domain. domain model generally uses the vocabulary of the domain, thus allowing a representation of the model to be communicated to non-technical stakeholders. It should not refer to any technical implementations such as databases or software components that are being designed.



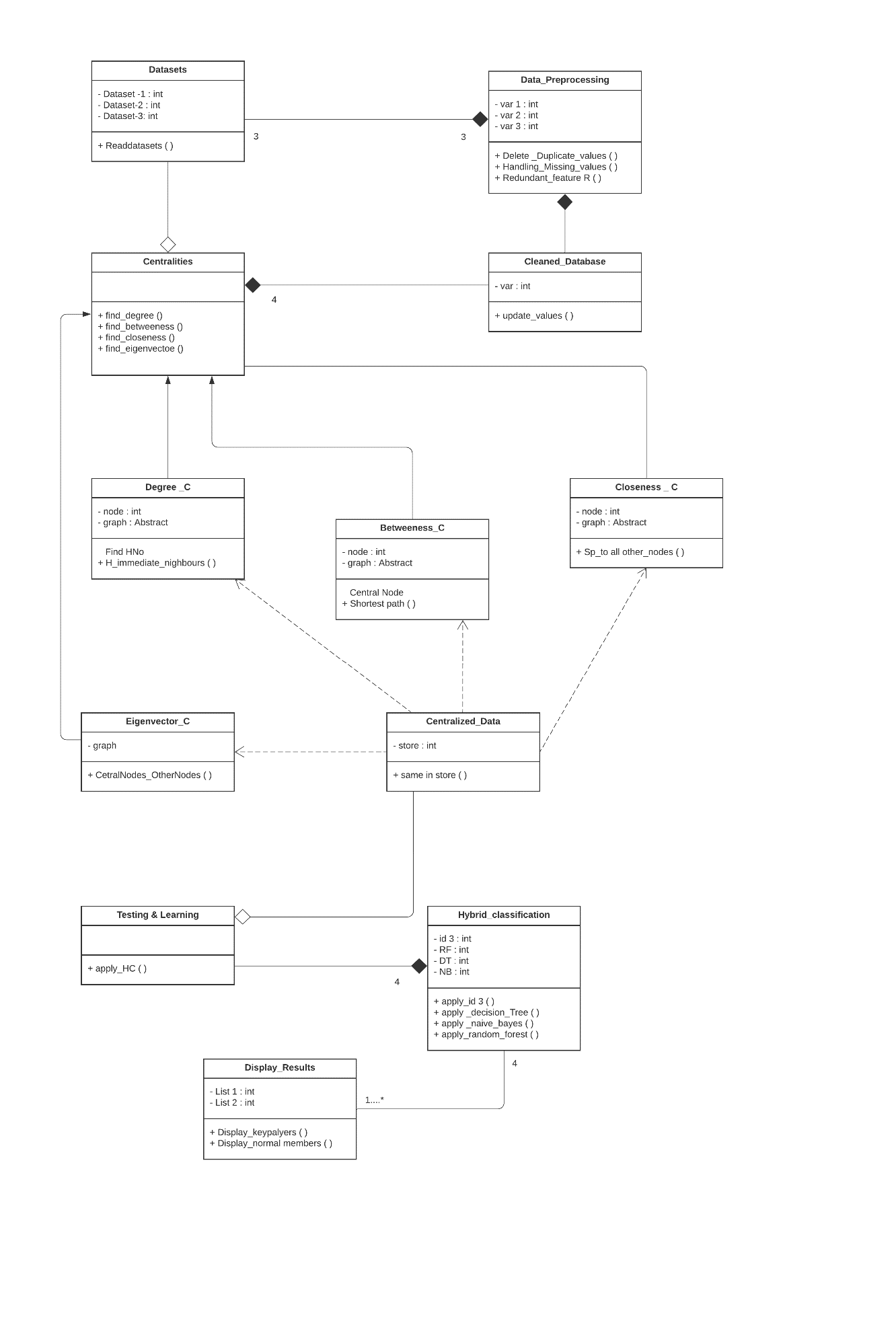
## Entity Relationship Diagram with data dictionary

ER model is commonly formed to represent things that a business needs to remember in order to perform business processes.



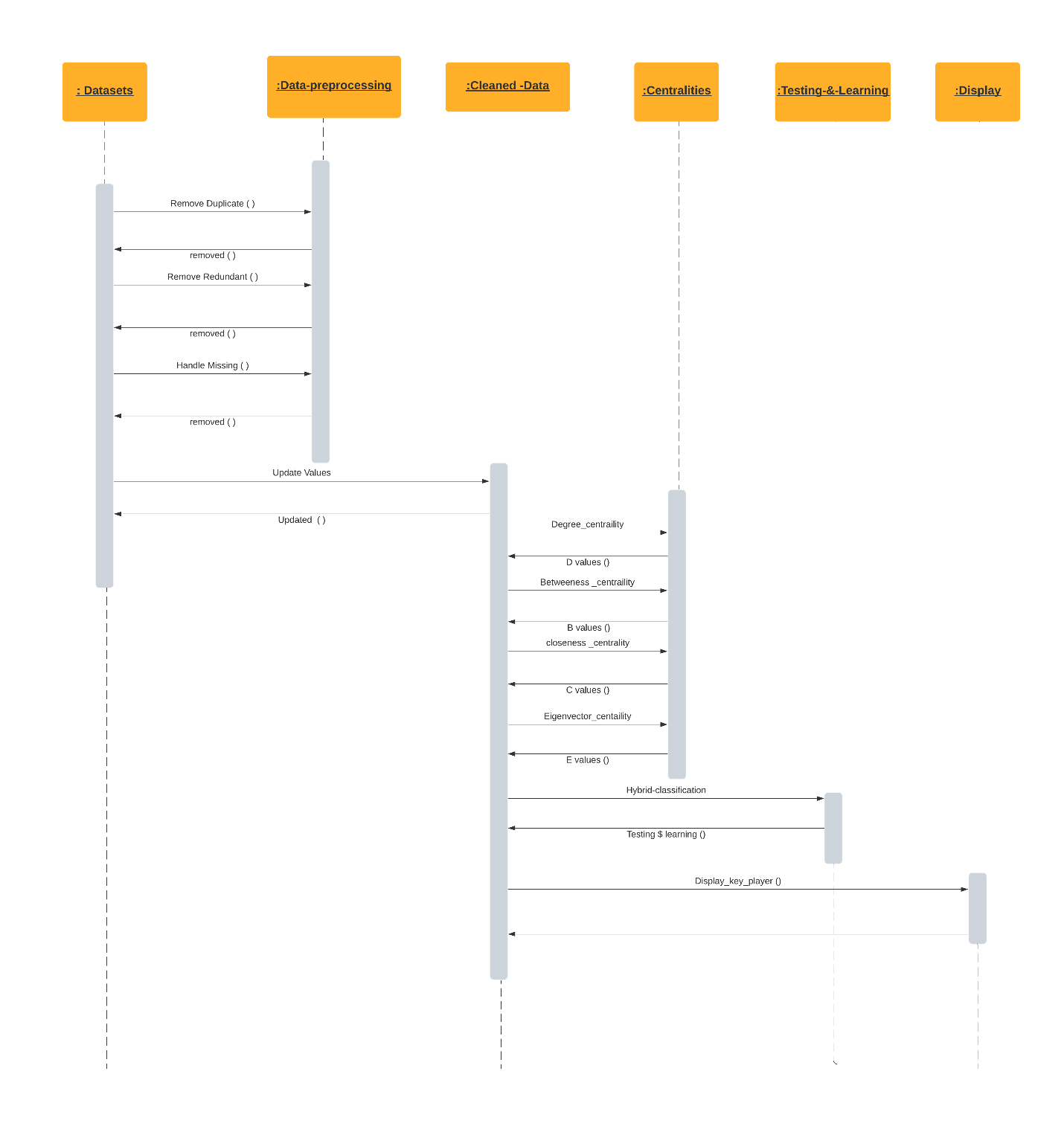
## Class Diagram

The class diagram is the main building block of [object-oriented](https://en.wikipedia.org/wiki/Object-oriented_programming) modeling. It is used for general [conceptual modeling](https://en.wikipedia.org/wiki/Conceptual_model) of the systematic of the system and for detailed modeling translating the models into [programming code](https://en.wikipedia.org/wiki/Programming_code). Class diagrams can also be used for [data modeling](https://en.wikipedia.org/wiki/Data_modeling). The classes in a class diagram represent both the main elements, interactions in the system and the classes to be programmed.

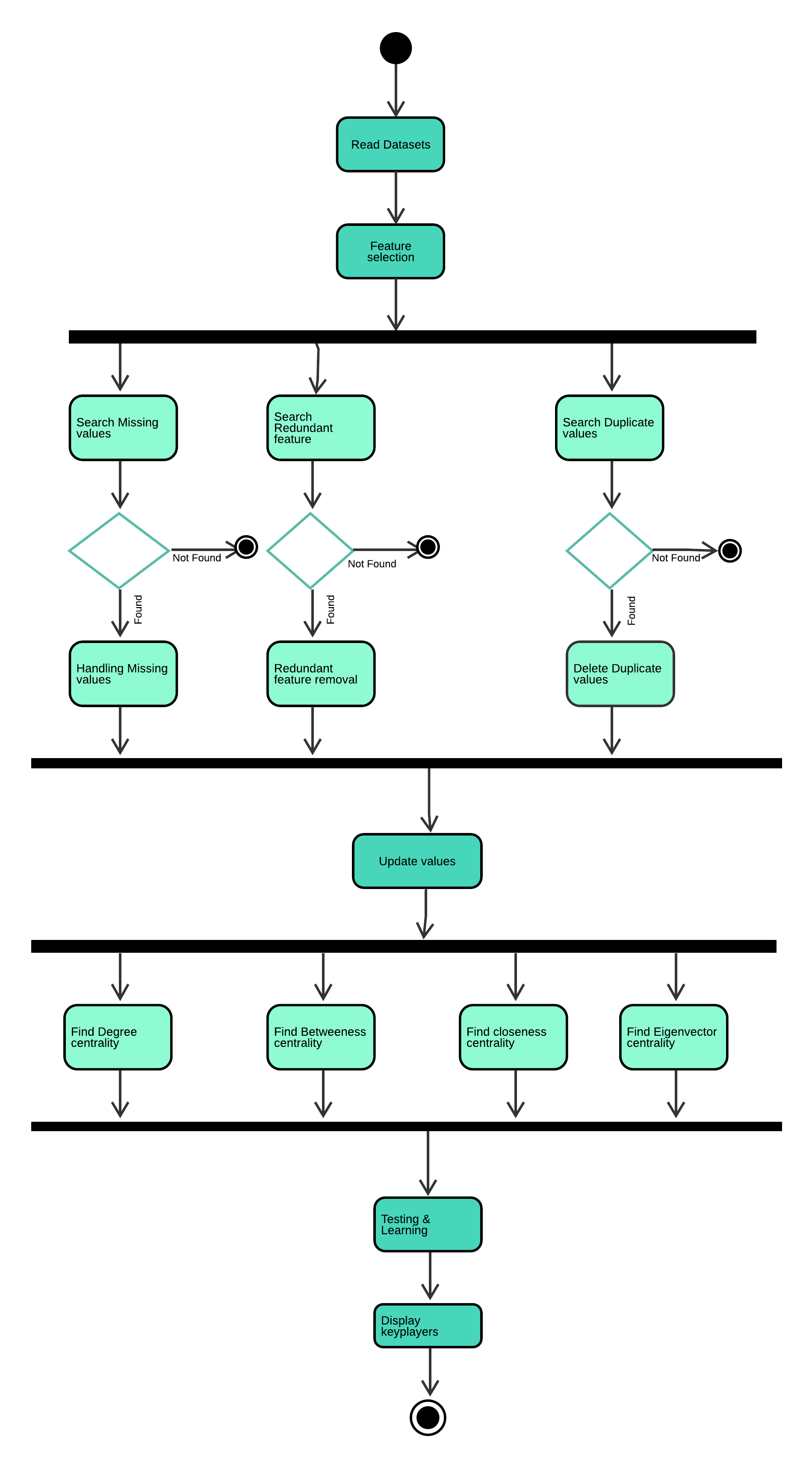


## Sequence / Collaboration Diagram

A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. Diagrams are typically associated with use case realizations in the Logical View of the system under development. Sequence diagrams are sometimes called **event diagrams** or **event scenarios**.

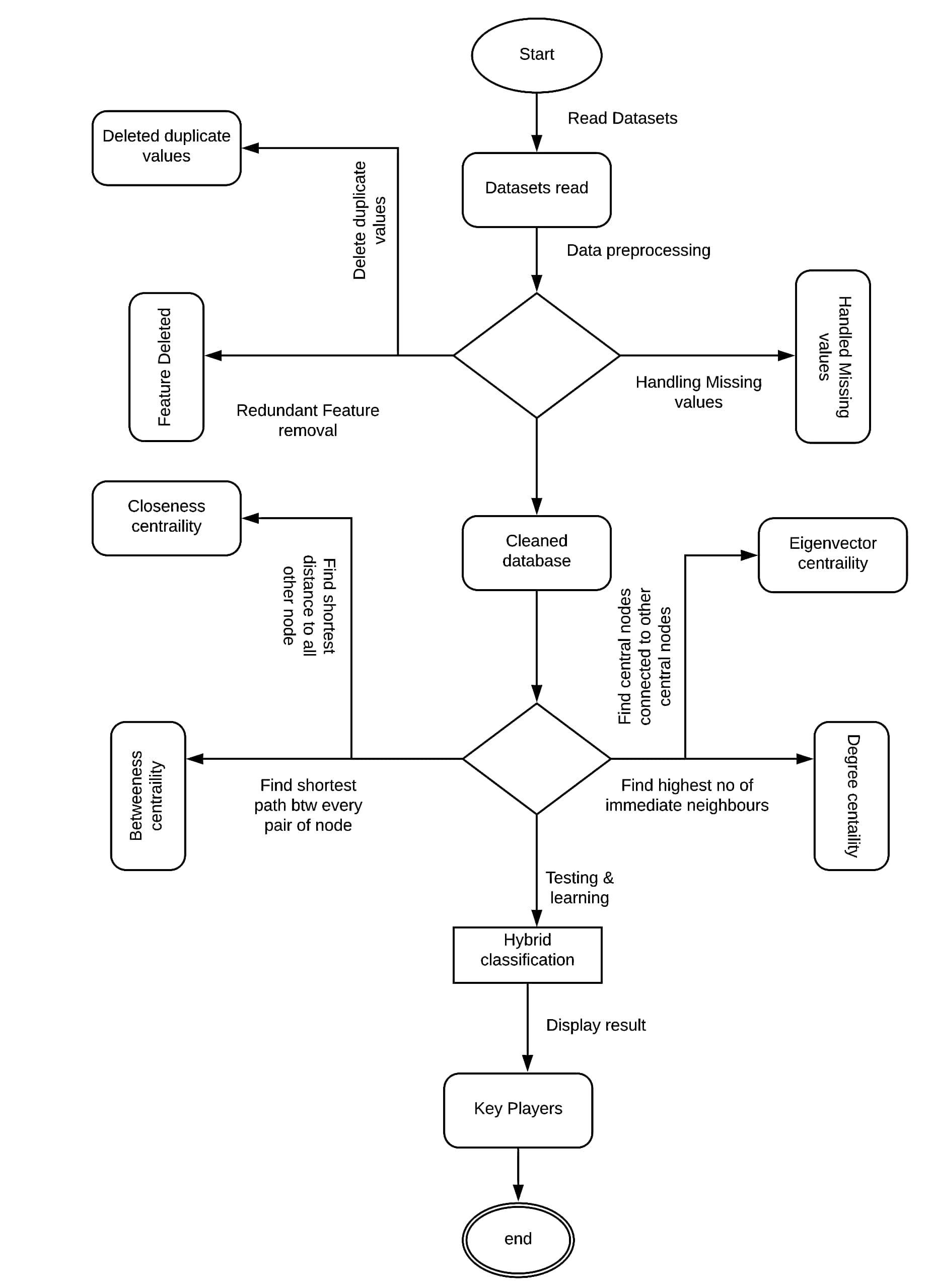


## Activity Diagram



## State Transition Diagram

A state diagram is a type of [diagram](https://en.wikipedia.org/wiki/Diagram) used in [computer science](https://en.wikipedia.org/wiki/Computer_science) and related fields to describe the behavior of systems. State diagrams require that the system described is composed of a finite number of [states](https://en.wikipedia.org/wiki/State_(computer_science)); sometimes, this is indeed the case, while at other times this is a reasonable [abstraction](https://en.wikipedia.org/wiki/Abstraction_(computer_science)). Many forms of state diagrams exist, which differ slightly and have different [semantics](https://en.wikipedia.org/wiki/Semantics#Computer_science).



## Data Flow diagram

A Data Flow Diagram (DFD) is a traditional way to visualize the information flows within a system. A neat and clear DFD can depict a good amount of the system requirements graphically. It can be manual, automated, or a combination of both. It shows how information enters and leaves the system, what changes the information and where information is stored.

