

**A Project Report on**

**“ANALYSIS OF FORIGN DIRECT INVESTMENT IN**

**INDIA CONSIDERING VARIOUS PARAMETERS”**

Submitted in partial fulfillment as a requirement for the award of degree of

**BACHELOR OF COMPUTER APPLICATION**

**Submitted by**

**KAVITHA P [U01BH21S0067]**

**LIKHITHA B K [U01BH21S0131]**

**SANDESH GOWDA B M[U01BH21S0132]**

**Under the guidance of**

**Mrs. NAGASHREE S**

Assistant Professor,

Dept. of Computer Application

DEPARTMENT OF COMPUTER APPLICATION

**SBRR MAHAJANA FIRST GRADE COLLEGE (Autonomous)**

JAYALAKSHMIPURAM,

MYSORE-12

College with Potential for Excellence

**EDUCATION TO EXCEL**

***SBRR MAHAJANA FIRST GRADE COLLEGE (Autonomous)***

Jayalakshmipuram, Mysore-12

(Affiliated to University of Mysore and Accredited by NAAC with ‘A’ Grade)

DEPARTMENT OF COMPUTER APPLICATION



**Certificate**

This is to certify that Ms.KAVITHA P, Ms.LIKHITHA B K , Mr. SANDESH GOWDA B L With Register number U01BH21S0067,U01BH21S0131,U01BH21S0132 has successfully completed the Project work in “**Analysis of foreign direct investment in India considering various parameters**”

Signature of the teacher in charge Head of the Department

1. EXAMINER

**DECLARATION**

Student of VI Semester BCA, **Department of Computer Applications, SBRR Mahajana**  **First Grade College (Autonomous), Mysore**, do hereby declare that the project titled

“ANALYSIS OF FOREIGN DIRECT INVESTMENT IN INDIA CONSIDERING VARIOUS

PARAMETERS”. Has been carried by me at University of Mysore during period of February 2024 to May 2024. This project report is submitted in partial requirement for the award of the degree of Bachelor of Computer Application (BCA) by the University of Mysore.

This is a bona fide work and the matter embodied in the report has not been submitted previously by anybody for the award of any degree/diploma to any other Institution or University.

Place: Mysore

Date:

**KAVITHA P**

**LIKHITHA B K**

**SANDESH GOWDA B L**

**ACKNOWLEDGEMENT**

The life of human beings is always inspired and guided by certain individuals and events. Accordingly, whatever I achieved in my project has been the fruit of advice and good will of my project mentors. I take this opportunity to mention the people who have contributed to success of my project greatly indebted to our Principal Prof. **Dr B R Jayakumari ma'am, SBRR Mahajana First Grade College and Mr. A Radhesh, Head of the Department of Computer Application,** for the facilities and support extended towards us.

Our heartfelt gratitude to my Parents, who have been my support system throughout this journey.

I consider it is a privilege and honor to work under, our guide **Mrs. Nagashree S, Assistant** **Professor** **Dept. Of Computer Application** for their invaluable support, guidance, and encouragement.

We also much indebted and grateful to the other teaching and non-teaching staff of our dept. which extended their unlimited moral support.

At last, we would like thank our parents and friends for providing encouragement and moral support without which this was not possible.

Place: Mysore

Date:

**Signature:**

**KAVITHA P**

**LIKHITHA B K**

**SANDESH GOWDA B L**

**ABSTRACT**

Foreign Direct Investment (FDI) plays a crucial role in the economic development of countries, including India. Understanding the factors influencing FDI inflows and predicting future trends can provide valuable insights for policymakers, investors, and stakeholders. In this project, we aim to analyze FDI trends in India across various sectors from 2008 to 2024 and predict FDI inflows for the year 2025 using regression algorithms. Additionally, we employ clustering techniques to differentiate high and low-profit sectors based on historical data.

The project utilizes datasets obtained from data.gov.in, covering sectors such as Metallurgical Industries, Power, Mining, Natural Gases, Coal Production, Boilers and Steam Generating Plants, Electrical Equipment, Computer Software and Hardware, Electronics, Telecommunication, Automobiles, Media, Sea Transport, Air Transport, and Chemicals. The datasets are preprocessed to handle missing values, outliers, and ensure consistency. Exploratory Data Analysis (EDA) is conducted to visualize trends, correlations, and patterns in the data. Feature engineering techniques are employed to extract relevant features that influence FDI inflows. Regression algorithms, including linear regression, decision trees, random forests, and gradient boosting, are trained on historical data to predict FDI for 2025. The performance of the regression models is evaluated using metrics such as Mean Squared Error (MSE), R-squared, and Mean Absolute Error (MAE), with hyper parameter tuning to optimize model performance.

Furthermore, K-means clustering is applied to segment the data based on profitability, identifying clusters representing high and low-profit sectors within the FDI dataset. The predictions for 2025 are made based on the trained regression models, considering the latest available data. Uncertainty and potential risks associated with the predictions are evaluated to provide robust insights. The results of the regression analysis and clustering provide valuable insights into the factors driving FDI inflows in India. The project concludes with recommendations for policymakers, investors, and stakeholders, based on the findings of the analysis. Overall, this project contributes to the understanding of FDI dynamics in India and informs strategic decision-making in the context of economic development and investment.

**TABLE OF CONTENTS**

**Contents: Page no:**

Chapter 1: Introduction

1. Introduction 01
   1. source of data collection 02
   2. problem statement 03

* 1. objectives 04
  2. scope of the project 04

Chapter 2: Literature review

2.1 Literature review 05

2.2 Datasets 07

2.3 Background

2.3.1Regression technique 08

2.3.2 K-means clustering 09

2.4 Accuracy measures 10

2.5 Technology used 10

Chapter 3: System Design

3.1 Introduction to design document 12

3.2 Input 24

3.3 Function requirements

3.3.1 Data maintenance 24

3.3.2 Implementation of algorithm

3.3.2.1 Regression algorithm 25

3.3.2.2 K-means clustering 25

3.3.3 Analysis 26

3.3.4 Visualization 27

Chapter 4: System Implementation

4.1 Introduction 28

4.2 Overview 28

4.2.1 Result based on parameters 29

4.2.2 Result based on K-means clustering 29

4.3 Algorithm to depict overall steps

4.3.1 Algorithm for regression techniques 29

4.3.2 K-means clustering 31

4.4 Pseudo code for algorithm used

4.4.1 Regression algorithm 32

4.4.2 Pseudo code K-means clustering 37

Chapter 5: Experimental result and analysis

5.1 Experimental setup 40

5.2 Result based on algorithms 40

5.2.1 Linear regression 41

Chapter 6: System Testing

6.1 Introduction 43

6.2 Objective of testing 43

6.3 Testing methods

6.3.1 Unit testing 44

6.3.2 Integration testing 44

6.3.3 Output testing 44

6.3.4 User acceptance testing 45

6.4 Validation 45

6.5 Test report 45

Chapter 7: Conclusion and scope for future enhancement

7.1 Conclusion 48

7.2 Scope for future enhancement 49

7.3 Screen shots

7.3.1 Home page 50

7.3.2 Prediction 50

7.3.3 Classification 51

7.3.4 Comparison 51

Bibliography 52

**CHAPTER 1**

**INTRODUCTION**

**1. INTRODUCTION**

Data analysis, also known as analysis of data or data analytics, is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making.

Foreign direct investment (FDI) in India is a major monetary source for [economic development in India.](https://en.wikipedia.org/wiki/Economic_development_in_India) Foreign companies [invest directly i](https://en.wikipedia.org/wiki/Foreign_direct_investment)n fast growing private Indian businesses to take benefits of cheaper wages and changing business environment of India. [Economic liberalization](https://en.wikipedia.org/wiki/Economic_liberalisation_in_India) started in India in wake of [the 1991 economic crisis a](https://en.wikipedia.org/wiki/1991_Indian_economic_crisis)nd since then FDI has steadily increased in India, which subsequently generated more than one crore jobs. According to the [*Financial Times*,](https://en.wikipedia.org/wiki/Financial_Times) in 2015 India overtook [China a](https://en.wikipedia.org/wiki/China)nd the [US a](https://en.wikipedia.org/wiki/US)s the top destination for the Foreign Direct Investment. In first half of the 2015, India attracted investment of $31 billion compared to $28 billion and $27 billion of China and the US respectively.

Data mining is a particular data analysis technique that focuses on modelling and knowledge discovery for predictive rather than purely descriptive purposes, while business intelligence covers data analysis that relies heavily on aggregation, focusing on business information. Data mining uses many different techniques and algorithms to discover the relationship in large amount of data. It is considered one of the most important tools in information technology in the previous decades.

Regression algorithm is method we use to analyze the FDI data and helps in predicting the future result based on the previous data sets. Similarly, the K-means clustering algorithm is use to find out which states are similar to each other considering FDI, and which states are safer have high growth rate and low growth rate, clustering algorithm was performed on the FDI dataset.

We used the FDI datasets for our study. The datasets are downloaded from the website data.gov.in. We collected data sets from 2008-2024. The analysis part involves the collecting the data and performing the statistics operation on the previous datasets and predicting the result of 2025.

1.1 NEED FOR THE PROJECT SOURCE OF DATA COLLECTON

In the development of the predictive model the data sets were collected internally in secondary form. Secondary data imply statistical materials or information not originated or obtained by the investigator himself, but obtain from someone’s record or published source such as the central government agencies.

DATA SOURCES:

<https://data.gov.in/major-indicator/foreign-direct-investment-fdi-agriculture>

|  |  |  |
| --- | --- | --- |
| Volume of data | No of tables | No of Records |
| 100MB | 1 | 16000 |

1.2 PROBLEM STATEMENT

Problem Statement

* There are some foreign Direct Investment which makes profit and loss to the Government.

* The Government cannot predict the future profit from the FDI

* One would think that acts of terrorism would have a negative impact on Foreign Direct Investment (FDI) flows to affected countries

Parameters

* Metallurgical industries
* Power
* Mining
* Natural gases
* Coal production
* Boilers and steam generating plants
* Electrical equipment
* Computer software and hardware
* Electronics
* Telecommunications
* Automobile
* Media
* Sea Transport
* Air Transport chemical

1.3 OBJECTIVES

* Prediction of results
* For 2025 for all the above parameters
* Comparing all year's data for different parameters.
* Helping the Gov. To analyze the low and high profit FDI
* Generating the year's wise report of all parameters
* Developing the Dashboard for all the above

1.4 SCOPE OF THE PROJECT

* The model help the foreign direct Investment control authorities concerned with similar assignment to assess their performance over the years.
* The model helps the FDI authorities or user of the investment on product help them to analyze the profit or loss for their investment.
* The model will awaken the sense of responsibility of FDI users and government.

**CHAPTER 2**

**LITERATURE REVIEW**

2.1 Literature Review

1. Foreign Direct Investment (FDI) is fund flow between countries by which one can gain some benefit from their investment, while another can enhance the productivity and find a better position through performance. The effectiveness and efficiency depends upon the investor’s perception: If an investment is long term, then it contributes positively towards the economy. If it is short term for the purpose of making profit, then its economic impact may be less significant. The FDI may also be affected by government trade barriers and policies for foreign investments and may be less or more effective towards contribution in economy and in GDP of the economy as well as the information technology sector. The success of Indian software industry has captured the imagination of both India and the world. For a country, attracting an inflow of FDI strengthens the connection to world trade networks and finances its development path. Foreign investment plays a significant role in development of India’s economy. FDI acts as a bridge to fill the gap between investment and saving. Many countries provide incentives for attracting the FDI. A foreign direct investor may be classified in any sector of the economy and could be any one – an individual, a group of related individuals, a public company or a private company, a government body or any combination of these.

1. India received highest foreign direct investment (FDI) in the world during the first half of 2025, leaving bigger economies like the US and China behind. In the process of globalization, India has liberalized all its sectors and invited FDI in most of the sectors, albeit with a sectorial cap. Internationalization of banks is perhaps the best example of India’s globalization. There are 44 foreign banks with 300 branches operating in India having a cap of 74 per cent and 20 per cent foreign investment in private and public sector banks, respectively. The present study aims to determine the motives behind bank FDI inflow into India. To accomplish that, a county wise panel was constructed and bank FDI data from 2008to 2024 was analyzed through generalized method of moments, a dynamic panel data model. The result of the study shows that bank FDI follows overall FDI, indicating that foreign banks follow their clients from their home country to serve them in the host country. However, powered advantages offer them profit-making opportunities and thus play a limited role in drawing bank FDI, which contribute to the development of the Indian economy. The argument that bank FDI inflow increases during a period of crisis is not relevant in the Indian context. The study suggests increasing the FDI cap in banking sector to attract more FDI and further relax the current restrictive policy on entry of foreign banks in India.

1. India has emerged as the most favored destination for foreign direct investment (FDI) in 2015 so far, outpacing China and the US, London-based business daily Financial Times (FT) said in a report on Tuesday. FDI inflows into India during January-June stood at $31 billion, ahead of China’s $28 billion and the US’s $27 billion, said the FT report under the headline

‘India grabs investment league pole position’.

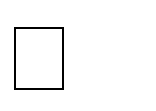
[4]The Author find that investments of the sort made by the Global Alliance for Vaccination and Immunization (Gavi) are linked to measurably higher DTP rates, and we estimate the returns to such investments as at least 12:1. For countries on the threshold of graduating from Gavi, these results suggest that much is at stake: reducing investment in vaccination could reverse both the public health improvements and tangible economic benefits associated with ongoing improvements in vaccination rates.

1. Interest in nominal GDP (NGDP) targeting has come in the context of large advanced economies. Developing countries are better suited for it, however, in light of big supply shocks and terms of trade shocks, such as monsoon rains and oil import price shocks in the case of India. Under annual inflation targeting (IT), the full impact of adverse supply shocks is felt as lost real GDP. NGDP targeting automatically accommodates such shocks, while retaining the advantage of anchoring expectations. We derive the condition under which NGDP targeting would dominate other regimes such as annual IT, to achieve objectives of output and price stability. We estimate key parameters for the case of India and conclude that the condition may indeed hold.

1. In business, consumer’s interest, behavior, product profits are the insights required to predict the future of business with the current data or historical data. These insights can be generated with the statistical techniques for the purpose of forecasting. The statistical techniques can be evaluated for the predictive model based on the requirements of the data. The prediction and forecasting are done widely with time series data. Most of the applications such as Metallurgy Industries forecasting, finance and stock market combine historical data with the current streaming data for better accuracy. However the time series data is analyzed with regression models. In this paper, linear regression and support vector regression model is compared using the training data set in order to use the correct model for better prediction and accuracy.

2.2 Datasets

A *dataset* (or *data set*) is a collection of data, usually presented in tabular form. Each column represents a particular variable. Each row corresponds to a given member of the *dataset* in question. It lists values for each of the variables, such as height and weight of an object. In the development of the predictive model the data sets were collected internally in secondary form. Secondary data imply statistical materials or information not originated or obtained by the investigator himself, but obtain from someone’s record or published source such as the central government agencies. There are many FDI investment in India like metallurgy industries, power, sea transport, Agriculture, power etc.

 The data sets are collected from the data.gov.in website. The previous year’s datasets from 20082024 were downloaded from website. There are total 10000 of datasets up to 100mb of datasets are collected.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Datasets | |  | Country | #transaction | Date | volume |
|  |  |  | India | 18 | 2008-2024 | 2mb |
| Metallurgical industries |
| Power | |  | India | 18 | 2008-2024 | 2mb |
| Mining | |  | India | 18 | 2008-2024 | 2mb |
| Natural Gas | |  | India | 18 | 2008-2024 | 2mb |
| Coal Production | |  | India | 18 | 2008-2024 | 2mb |
| Boiler and stream | |  | India | 18 | 2008-2024 | 2mb |
| Electronics | |  | India | 18 | 2008-2024 | 2mb |
| Mobile | |  | India | 18 | 2008-2024 | 2mb |
| Media | |  | India | 18 | 2008-2024 | 2mb |
| Transport | |  | India | 18 | 2008-2024 | 2mb |

2.3 Background

**2.3.1 Regression Technique**

In [statistics, **l**](https://en.wikipedia.org/wiki/Statistics)**inear regression** is a [linear ap](https://en.wikipedia.org/wiki/Linear)proach for modelling the relationship between a scalar [dependent variable *y*](https://en.wikipedia.org/wiki/Dependent_variable) and one or more [explanatory variables (o](https://en.wikipedia.org/wiki/Explanatory_variable)r independent variables) denoted *X*. The case of one explanatory variable is called [simple linear regression. F](https://en.wikipedia.org/wiki/Simple_linear_regression)or more than one explanatory variable, the process is called multiple linear regression. In regression problem the goal of the algorithm is to predict real valued output. The main terminologies in

Regression Techniques are

* **Independent Variables (features):** An independent variable is a variable that is manipulated to determine the value of a dependent variable. Simply, they are The features which we want to use to predict some given value of Y. It can be also called an explanatory variable
* **Dependent Variable (target):** The dependent variable depends on the values of the independent variable. Simply put, it is the feature which we are trying to predict. This can also be commonly known as a response variable.

**2.3.2 K-means clustering**

**K-means clustering** is a method of vector quantization[, o](https://en.wikipedia.org/wiki/Vector_quantization)riginally from [signal processing, t](https://en.wikipedia.org/wiki/Signal_processing)hat is popular for cluster analys[is in](https://en.wikipedia.org/wiki/Cluster_analysis) data mini[ng. *k*](https://en.wikipedia.org/wiki/Data_mining)-means clustering aims to [partition *n*](https://en.wikipedia.org/wiki/Partition_of_a_set) observations into *k* clusters in which each observation belongs to the [cluster wi](https://en.wikipedia.org/wiki/Cluster_(statistics))th the nearest [mean, s](https://en.wikipedia.org/wiki/Mean)erving as a prototype of the cluster.

The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different power causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter’s of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop, we may notice that the k centroids change their power step by step until no more changes are done.

# 2.4 Accuracy Measures

Accuracy measures involves comparing each algorithm’s result with the actual result. The result for each algorithm is represented in tabular format ordered by their rank. The actual result is obtained by FDI per Parameter and representing them in the format similar to algorithmic result.

# 2.5 Technology used

**Net Beans**:

It is an integrated development environmen[t (I](https://en.wikipedia.org/wiki/Integrated_development_environment)DE) for [Java. N](https://en.wikipedia.org/wiki/Java_(programming_language))et Beans allows applications to be developed from a set of modular software component[s ca](https://en.wikipedia.org/wiki/Software_component)lled *modules*. Net Beans runs on [Microsoft Windows](https://en.wikipedia.org/wiki/Microsoft_Windows)[, MAC OS](https://en.wikipedia.org/wiki/MacOS)[, Linux an](https://en.wikipedia.org/wiki/Linux)d [Solaris. I](https://en.wikipedia.org/wiki/Solaris_(operating_system))n addition to Java development, it has extensions for other languages like [PH](https://en.wikipedia.org/wiki/PHP)[P, C](https://en.wikipedia.org/wiki/C_(programming_language))[, C+](https://en.wikipedia.org/wiki/C%2B%2B)[+, HTML5](https://en.wikipedia.org/wiki/HTML5) [, Javadoc an](https://en.wikipedia.org/wiki/Javadoc)d [JavaScript. A](https://en.wikipedia.org/wiki/JavaScript)pplications based on Net Beans, including the Net Beans IDE, can be extended by [third party developers.](https://en.wikipedia.org/wiki/Third_party_developer)

JAVA Servlets:

Java Servlets are server-side Java program modules that process and answer client requests and implement the servlet interface. It helps in enhancing Web server functionality with minimal overhead, maintenance and support. A servlet acts as an intermediary between the client and the server. As servlet modules run on the server, they can receive and respond to requests made by the client. Request and response objects of the servlet offer a convenient way to handle HTTP requests and send text data back to the client. Since a servlet is integrated with the Java language, it also possesses all the Java features such as high portability, platform independence, and security and Java database connectivity.

JSP (Java Server Pages):

Java Server Pages (JSP) is a technology for developing Webpages that supports dynamic content. This helps developers insert java code in HTML pages by making use of special JSP tags, most of which start with <% and end with %>. A Java Server Pages component is a type of Java servlet that is designed to fulfil the role of a user interface for a Java web application. Web developers write JSPs as text files that combine HTML or XHTML code, XML elements, and embedded JSP actions and commands. Using JSP, you can collect input from users through Webpage forms, present records from a database or another source, and create Webpages dynamically.

High charts:

**High charts** is a pure JavaScript based charting library meant to enhance web applications by adding interactive charting capability. High charts provides a wide variety of charts. For example, line charts, spline charts, area charts, bar charts, pie charts and so on.

MySQL:

MySQL is an open source relational database management system ([RDBMS) b](https://searchsqlserver.techtarget.com/definition/relational-database-management-system)ased on

Structured Query Language ([SQL).](https://searchsqlserver.techtarget.com/definition/SQL) MySQL runs on virtually all platforms, including

[Linux](https://searchdatacenter.techtarget.com/definition/Linux-operating-system)[, UNIX, a](https://searchdatacenter.techtarget.com/definition/Unix)nd [Windows. A](https://searchwindowsserver.techtarget.com/definition/Windows)lthough it can be used in a wide range of applications, MySQL is most often associated with web-based applications and online publishing and is an important component of an open sourc[e en](https://whatis.techtarget.com/definition/open-source)terprise stack called [LAMP. L](https://whatis.techtarget.com/definition/LAMP-Linux-Apache-MySQL-PHP)AMP is a Web development platform that uses [Linux as](https://searchdatacenter.techtarget.com/definition/Linux-operating-system) the operating system, [Apache as](https://searchmicroservices.techtarget.com/definition/Apache) the Web server, MySQL as the relational database management system and [PHP as](https://whatis.techtarget.com/definition/PHP-Hypertext-Preprocessor) the object-oriented scripting language.

**CHAPTER 3**

**SYSYTEM DESIGN**

# 3.1 Introduction to Design document

The Software Design will be used to aid in software development for android application by providing the details for how the application should built. Within the Software Design, specifications are narrative and graphical documentation of the software design for the project includes use case models, sequence diagrams and other supporting requirement information.

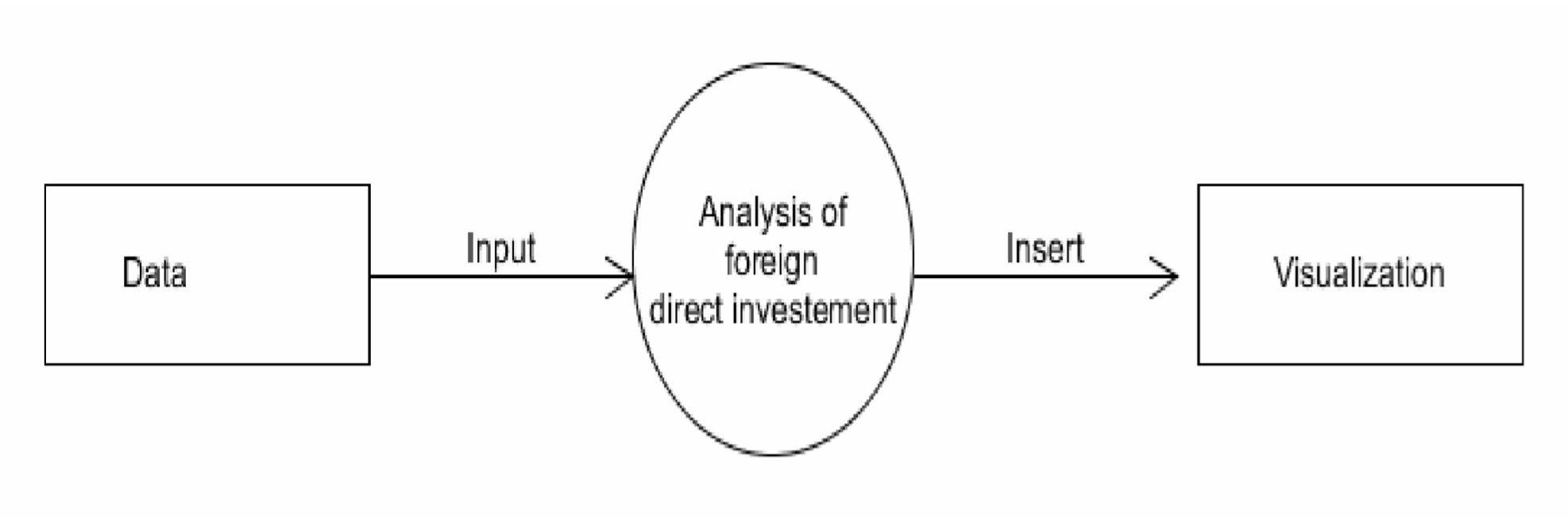
Scope

This software Design Document is for a base level system, which will work as a proof of concept for the use of building a system that provides a base level of functionality to show feasibility for large-scale production use. The software Design Document, the focus placed on generation of the documents and modification of the documents. The system will used in conjunction with other pre-existing systems and will consist largely of a document interaction faced that abstracts document interactions and handling of the document objects. This Document provides the Design specifications of FDI.

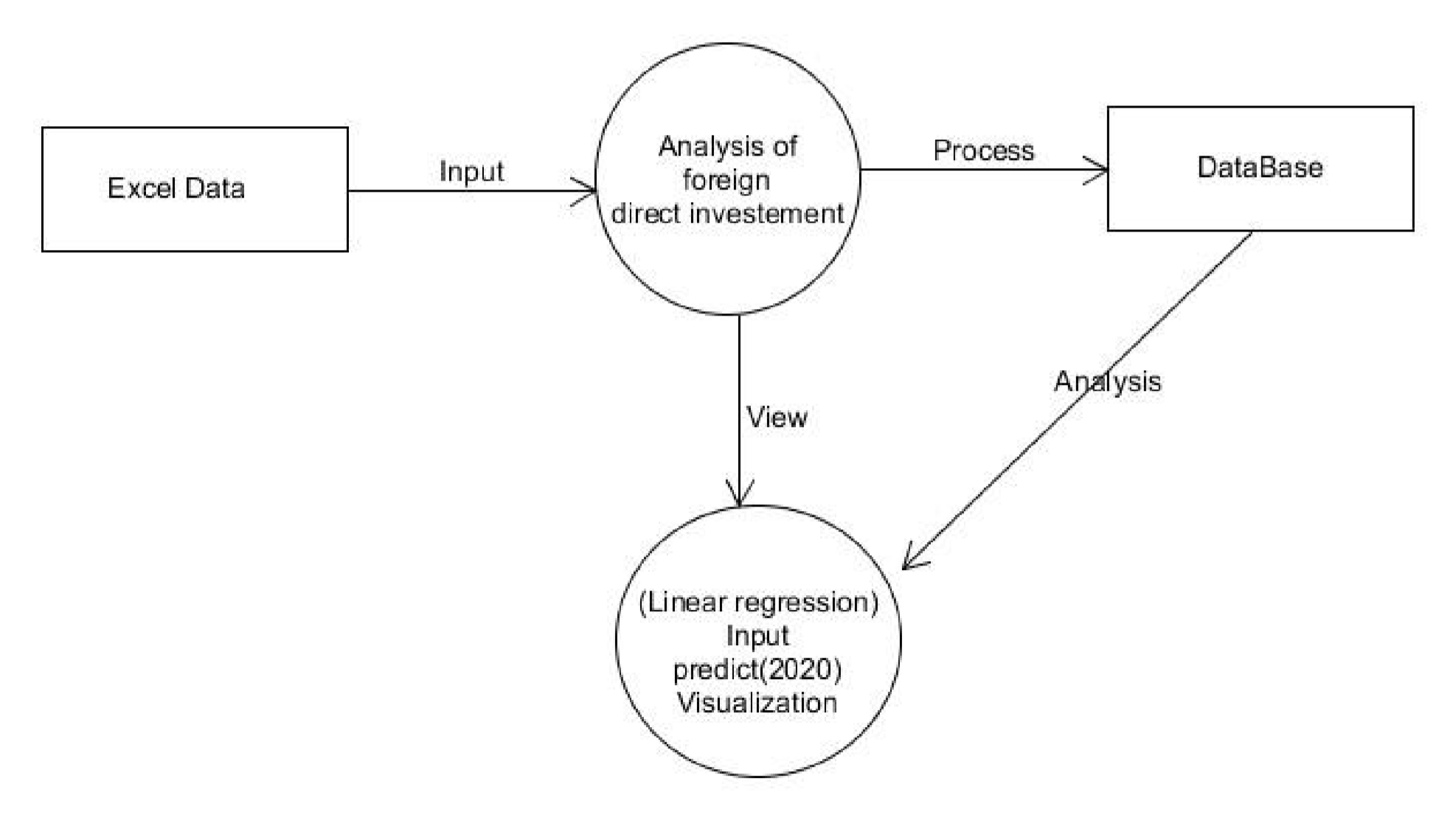
Data Flow Diagram

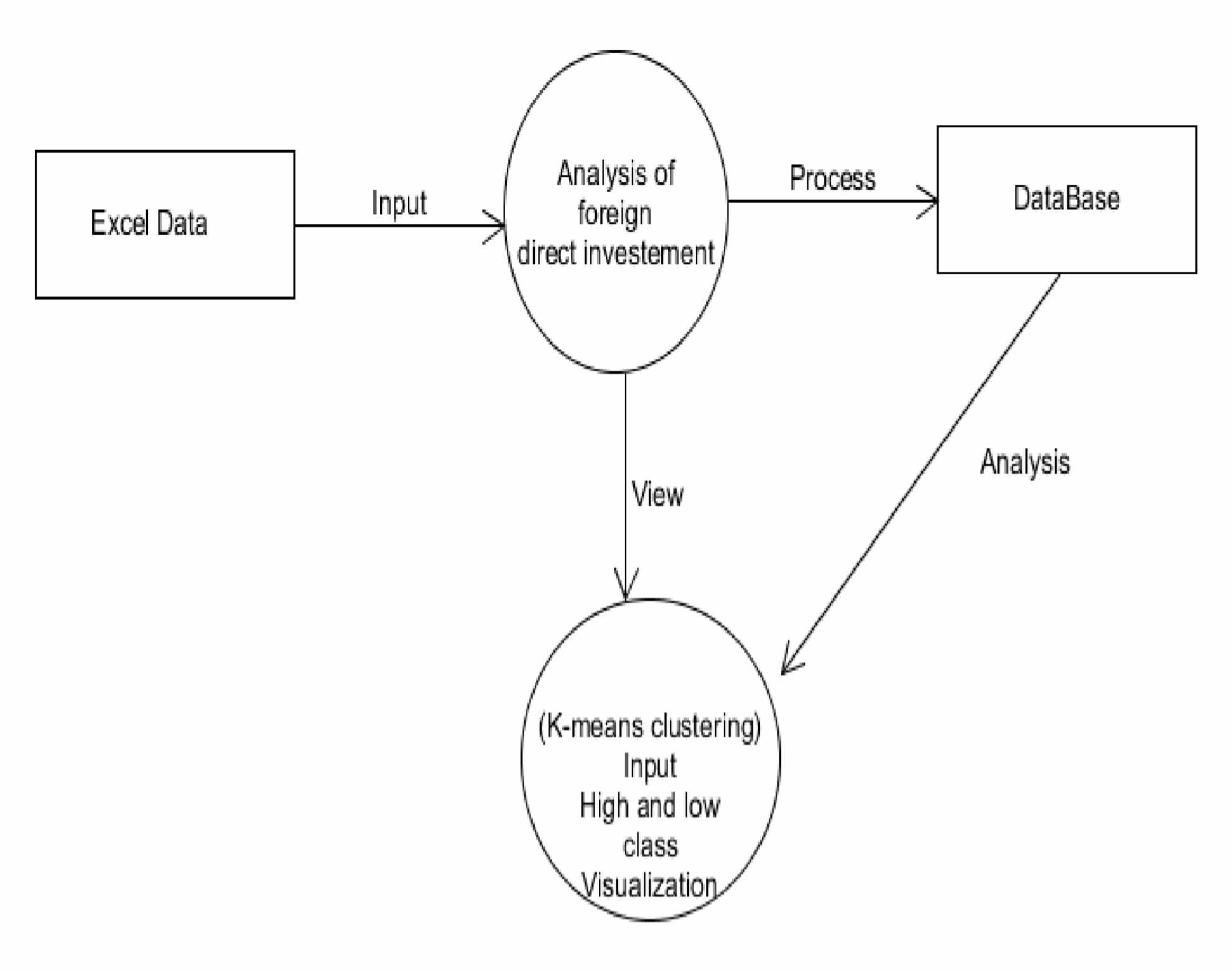
The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system

Level 1



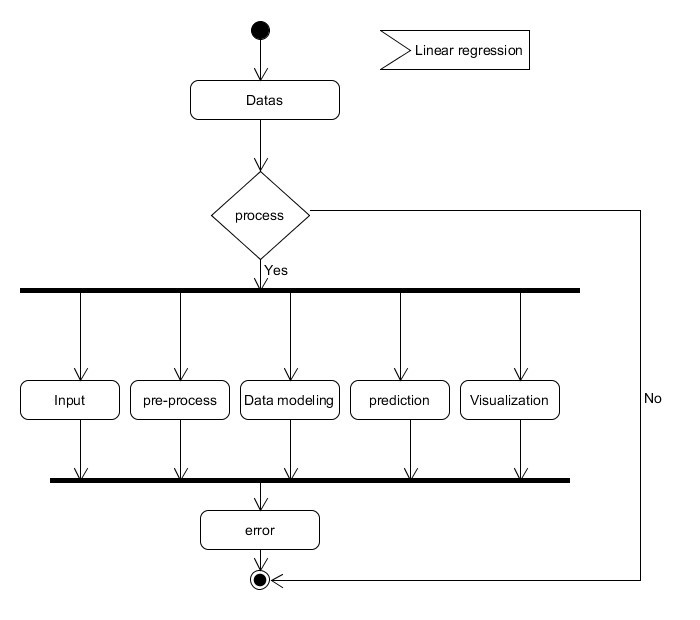
Level 2

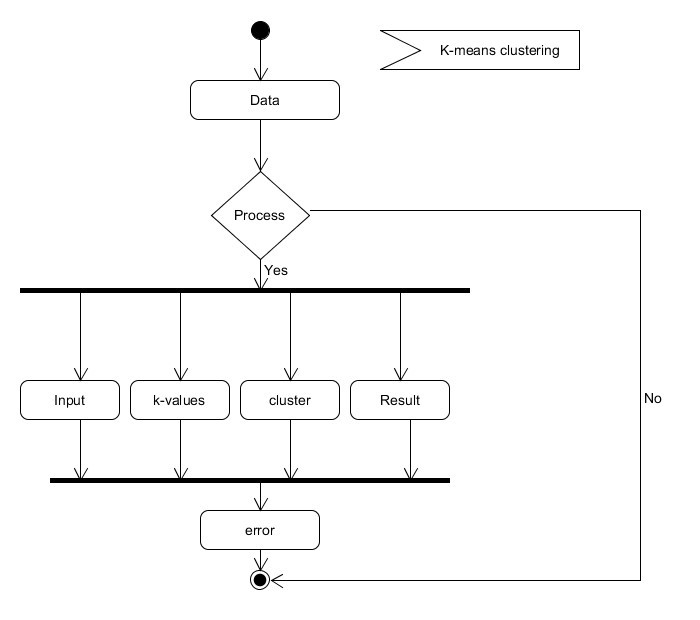




Activity Diagram

Activity diagrams represent the business and operational workflows of a system. An activity diagram is a dynamic diagram that shows the activity and the event that causes the object to be in the particular state. It is a simple and intuitive illustration of what happens in a workflow, what activities can be done in parallel, and whether there are alternative paths through the workflow.



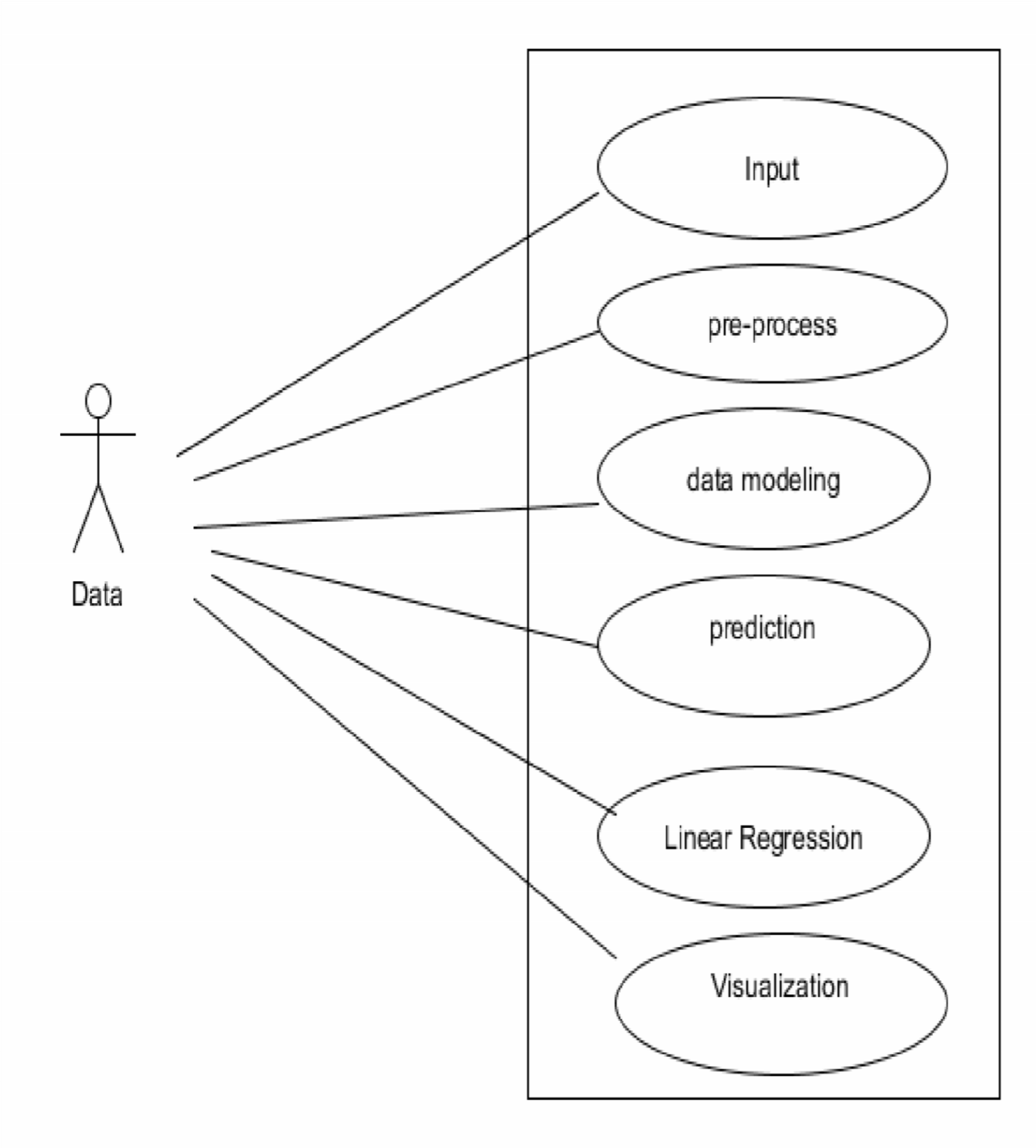


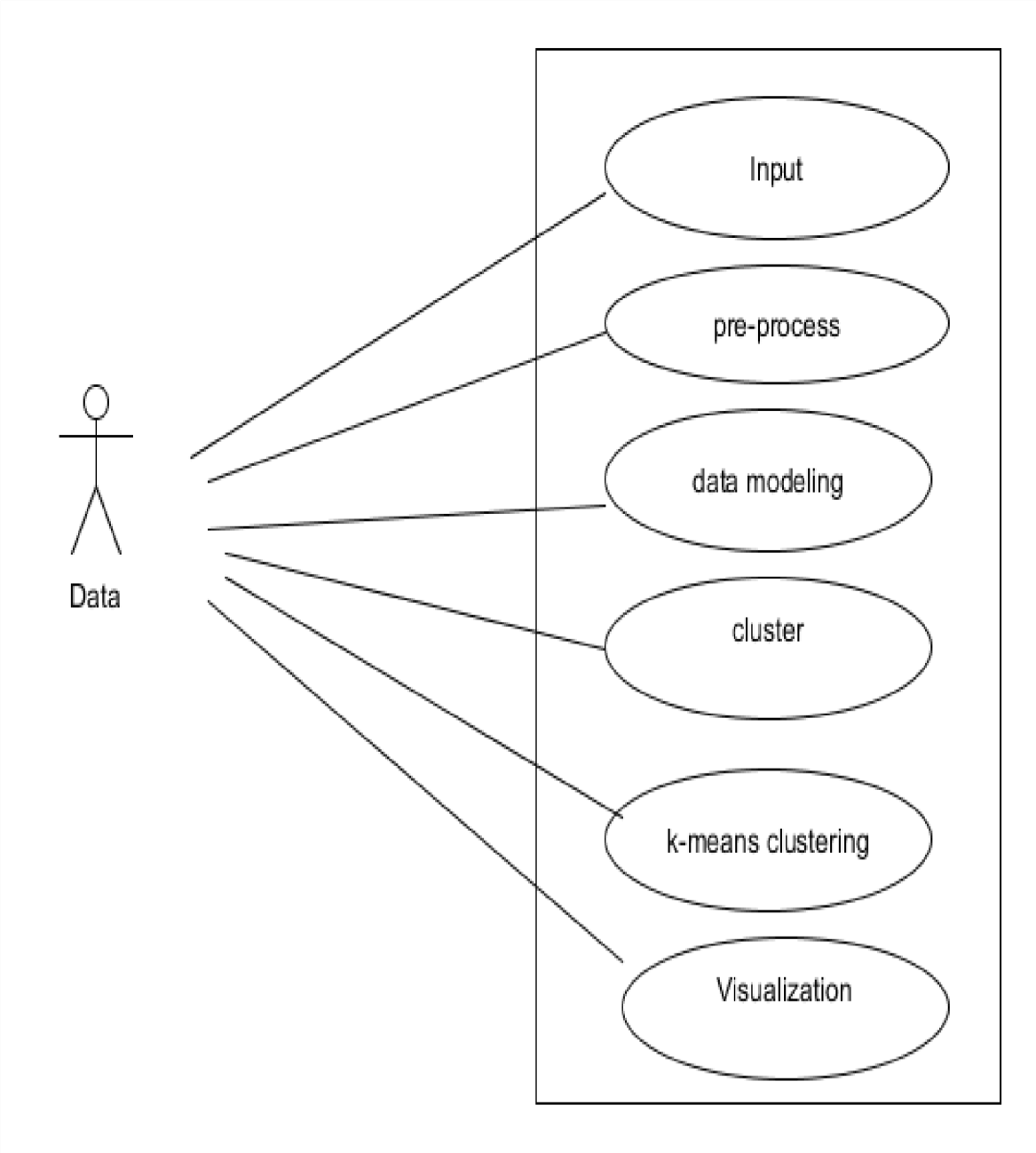
Use Case Diagram

Use case diagrams are valuable for visualizing the functional requirements of a system that will translate into design choices and development priorities.

They also help identify any internal or external factors that may influence the system and should be taken into consideration.

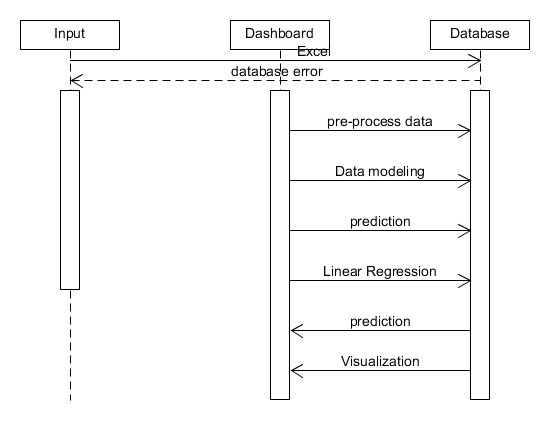
They provide a good high level analysis from outside the system. Use case diagrams specify how the system interacts with actors without worrying about the details of how that is implemented.

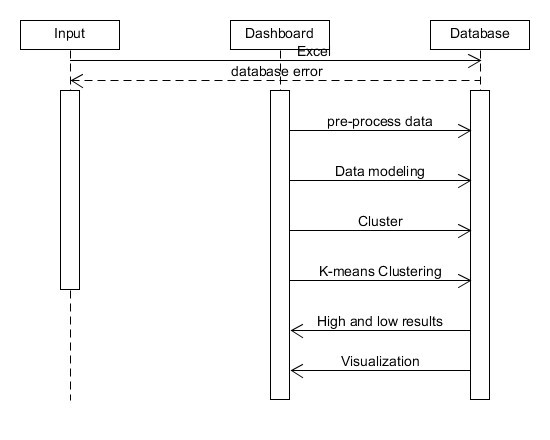




Sequence Diagram

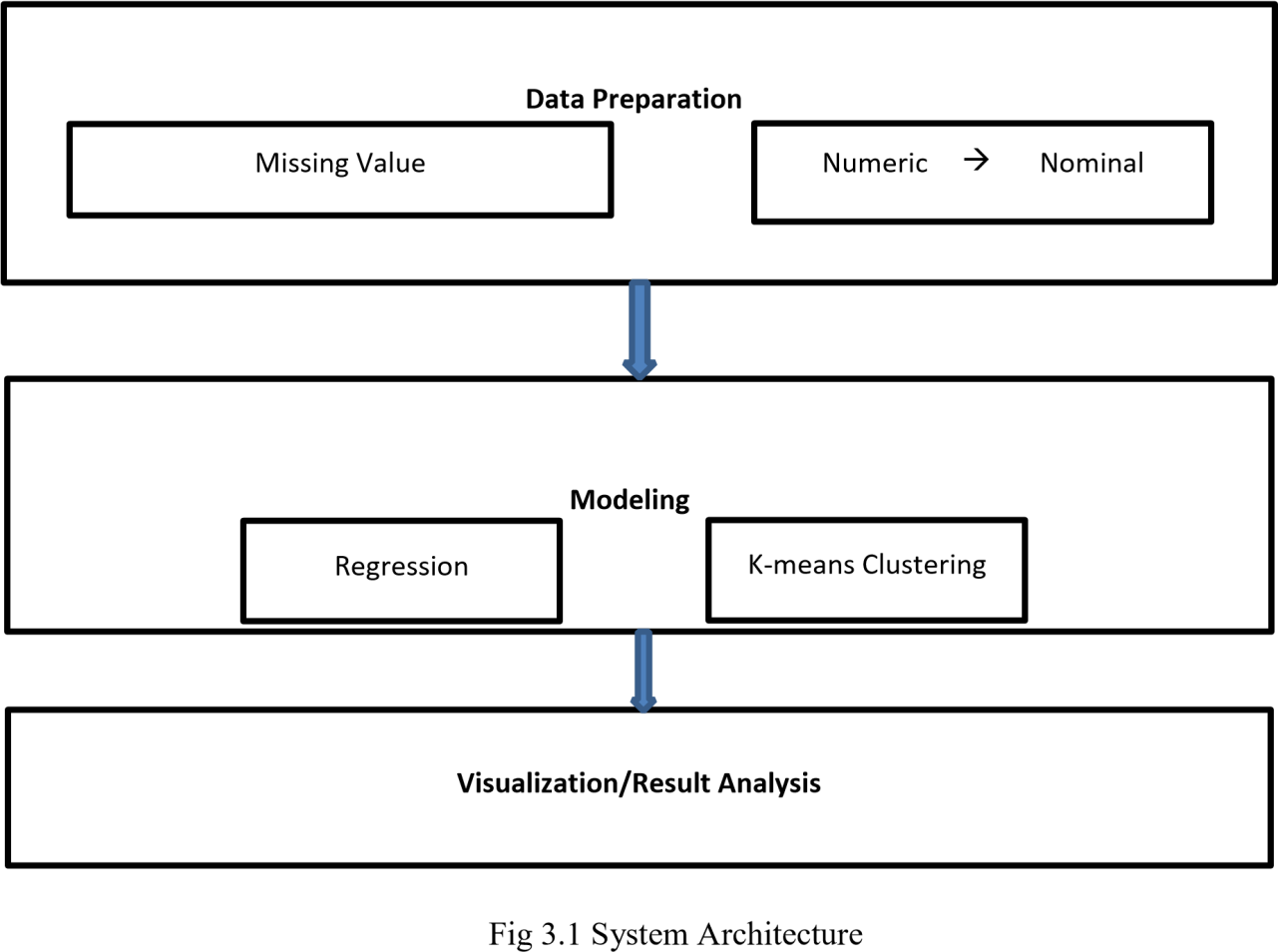
A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. Sequence diagrams are typically associated with use case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





Work Flow Diagram

The approach we took for our study follows the traditional data analysis steps



DATA PREPARATION

Data preparation was performed before each model construction. All records with missing value (usually represented by 0 in the dataset) in the chosen attributes were removed. All numerical values were converted to nominal value according to the data dictionary.

Missing Values: 

Occurs when the no data value is stored for the observation.

Example:

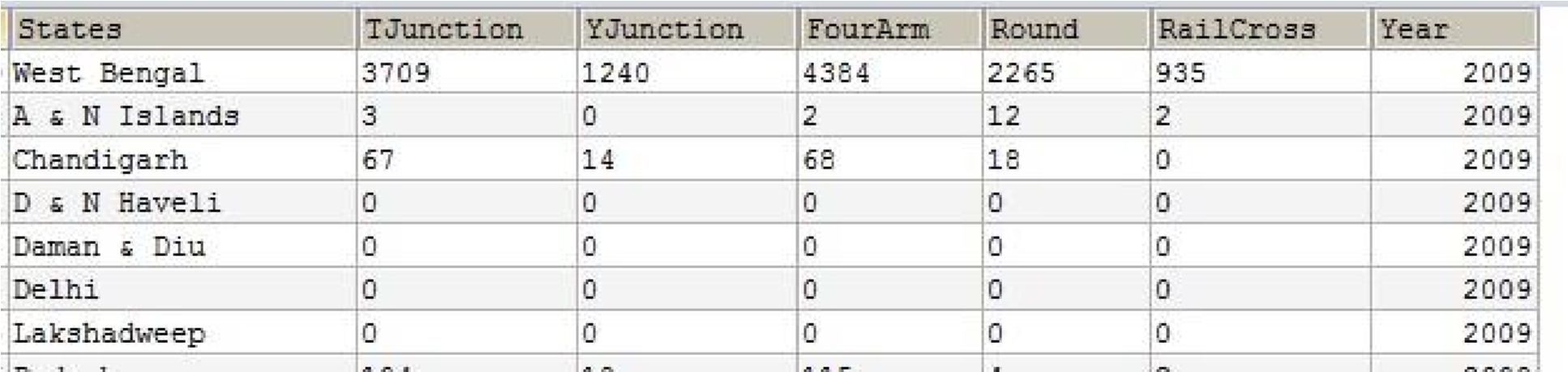


Fig 3.1 Missing Value

**Modeling**

We first calculate several statistics from the dataset to show the basic characteristics of the fatal s, then applied Regression and clustering relationships among the attributes and the patterns.

# 3.2 Input

* Metallurgical industries
* Power
* Mining
* Natural gases
* Coal production
* Boilers and steam generating plants
* Electrical equipment
* Computer software and hardware
* Electronics
* Telecommunications
* Automobile
* Media
* Sea Transport
* Air Transport chemical

# 3.3 Functional Requirements

###### 3.3.1 Data Maintenance

Data base service namely collecting processing, filtering, sorting, storing retrieving issuing transmitting and updating database entries.

The data that has been collected from the gov.in website is the cleaned data that stored in the CSV format

The stored CSV data are imported to the MySQL with help of JSON (Java Script Object Notation).

###### 3.3.2 Implementation of Algorithm

3.3.2.1 Regression Algorithm

A regression algorithm is designed to find the historical relationship between an independent and a dependent variable to predict the future values of the dependent variable. A regression models the past relationship between variables to predict their future behavior. The Algorithm uses the linear regression techniques based on the data set collected for the project. The linear regression technique helps in predicting the future behavior of FDI with help of the statistical methods. The algorithm finds the mean and variance value of the dependent variables, and apply the formula Y=b0+b1\*x to predict the future behavior.

Steps Involved:

1. Scan the transaction database and perform the operation on the missing values.
2. Calculate Mean and Variance.
3. Calculate Covariance.
4. Estimate Coefficients.
5. Make Predictions.
6. Predict Insurance.

3.3.2.2 K-means Clustering

K-means clustering algorithm was used to investigate the high and low-frequency powers. The algorithm follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main aim is to define k centroids, one for each cluster. M These centroids should be placed in a cunning way because of different power causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point re-calculate k new centroids as bary centers of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop, we may notice that the k centroids change their power step by step until no more changes are done.

Steps Involved:

1. Place K points into the space represented by the objects that are being clustered.

These points represent initial group centroids.

1. Assign each object to the group that has the closest centroid.
2. When all objects have been assigned, recalculate the positions of the K centroids.
3. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

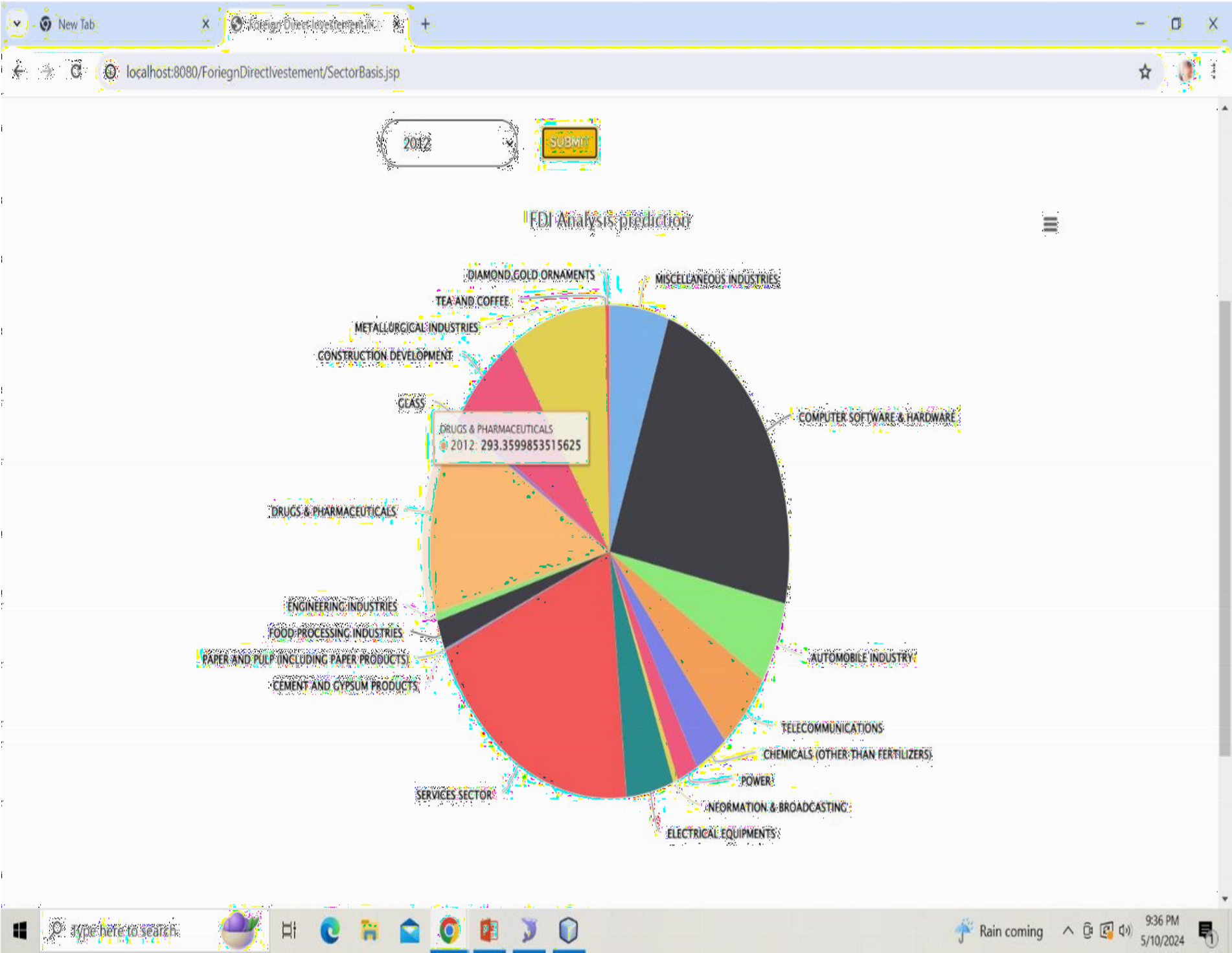
###### 3.3.3 Analysis

* The Clean Data was stored in the CSV format, the CSV data is imported into MySQL with help of JSON. Now data is ready to analyze by the data analyzing tools.
* Analysis of FDI involves like metallurgy industry, power, transport, sea transport, Agriculture, media etc.

Based on Analysis of 2001-2017 data sets each of the above parameters are analyzed using the Appropriate Algorithm

###### 3.3.4 VISUALIZATION

* Prediction of Dashboard for all the above.
* The year wise datasets from 2008-2024 are included in the prediction dashboard as bar graphs, pie charts, etc.
* The dashboard is more use full for the respective authorities that help them to know about the goings to happen in the current years, so they can take some preventive measures.
* The Application should produce the yearly report about the FDI.
* Year wise comparison of FDI.



**CHAPTER 4**

**SYSTEM IMPLEMENTATION**

# 4.1 Introduction

The project is implemented using java which is an object-oriented programming language. Object oriented programming is an approach that provides a way of modularizing program by creating partitioned memory area of both data and function that can be used as a template for creating copies of such module on demand.

This project is implemented using java programming language. Both servlet and JSP technologies are used to create a web application. Servlet are java programs are precompiled which can create dynamic web contents. There are many interfaces and class in the servlet API such as HTTP servlet, servlet request, servlet response etc. JSP is used to create a web application just as servlet.it can be thought of as an extension to servlet because it provides more functionality than servlet. MySQL server is used as a backend.

# 4.2 Overview

A regression algorithm is designed to find the historical relationship between an independent and a dependent variable to predict the future values of the dependent variable. A regression models the past relationship between variables to predict their future behavior.

K-means clustering algorithm was used to investigate the high and low-frequency powers. Further, they have been used association rule mining to recognize the association between the various factors related to FDI traffic s at various places with changeable occurrences

##### 4.2.1 Result based on parameters

The algorithm is implemented to fetch the result based on the parameters such as power, media, transport, agriculture etc.

The regression algorithm helps in predicting the future behavior of FDI considering various parameters. The dashboard provides the selection of states and type, based on the selection, the algorithm applies the statistical method to the previous data sets (2008-2024).

And predict the result of 2025.

##### 4.2.2 Result based on k-means clustering

K-means clustering algorithm was used to investigate the high and low-frequency rates. The centroid of clustering was calculated based on the total number of s happens to the total number of states. Based on the average value the cluster of high states and low are displayed separately.

# 4.3 Algorithm to depict overall steps:

##### 4.3.1 Algorithm for Regression Techniques

Input: FDI data set from 2008-2024.

**Parameters**

* Metallurgical industries
* Power
* Mining
* Natural gases
* Coal production
* Boilers and steam generating plants
* Electrical equipment
* Computer software and hardware
* Electronics
* Telecommunications
* Automobile
* Media
* Sea Transport
* Air Transport chemical

Output: Predicting the Result of 2025 shown with Bar graph.

Procedure:

Step 1: scan the transaction dataset.

Step 2: Handling missing data (all records with missing value is represented as 0 in the dataset)

Step 3: estimate the mean and the variance of both the input and output variables from the training data.

Step 4: Store the mean and variance result of x -axis.

Step 5: Similarly store the result of y-axis.

Step 6: Calculate the covariance with taking the mean result of x and y axis.

Step 7: Calculate coefficient value of b1 and b0 using the covariance.

Step 8: Once the coefficient is estimated, make the prediction Y=b0+b1(x).

##### 4.3.2 K-means Clustering

Input: FDI data set from 2008-2024

**Parameters**

* Metallurgical industries
* Power
* Mining
* Natural gases
* Coal production
* Boilers and steam generating plants
* Electrical equipment
* Computer software and hardware
* Electronics
* Telecommunications
* Automobile
* Media
* Sea Transport
* Air Transport chemical

Output: cluster of high and low frequency according to state wise.

Procedure:

Step1: scan transaction data set.

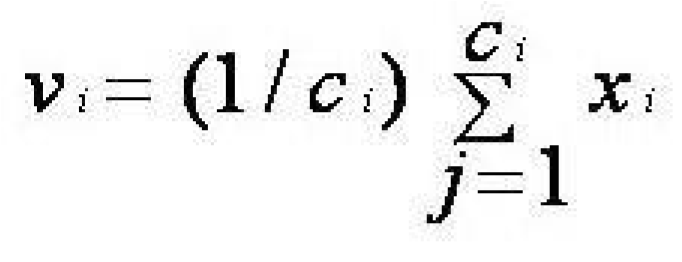
Step 2: calculate the average value based on the total number of s happens to the total number of states.

Step 3: assume that average value is centroid ‘c’, select ‘C’ cluster center.

Step 4: Calculate the distance between each data point and cluster centers.

Step 5: Assign the data point to the cluster center whose distance from the cluster Centre is minimum of all the cluster centers.

Step 6: Recalculate the new cluster Centre using:



Where, *‘ci’* represents the number of data points in *ith* cluster.

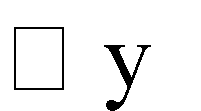
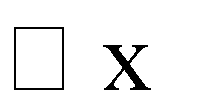
Step 7: Recalculate the distance between each data point and new obtained cluster centers.

Step 8: If no data point was reassigned then stop, otherwise repeat from step 5).

# 4.4 Pseudo code for Algorithm Used

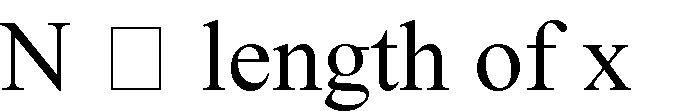
##### 4.4.1 Regression Algorithm

1. [] -axis values



1. [] -axis values

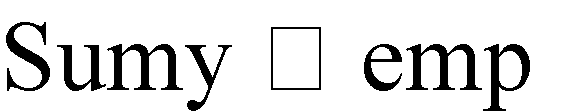
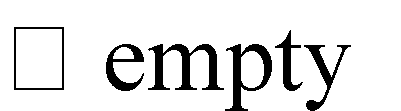
//Initialization

-axis

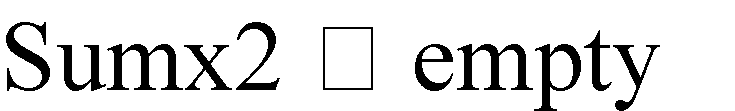
//First pass

//Initialization

Sumx



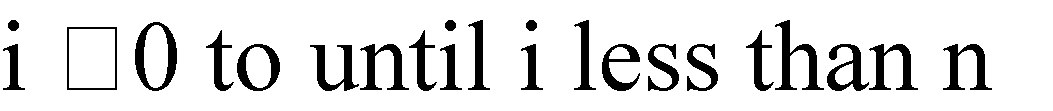
ty

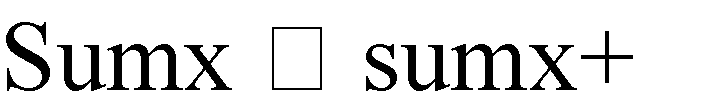


// first pass

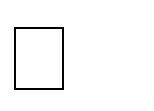
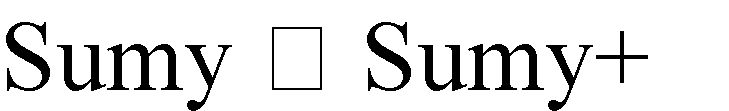
For loop Begins

Initialize



x []

y []



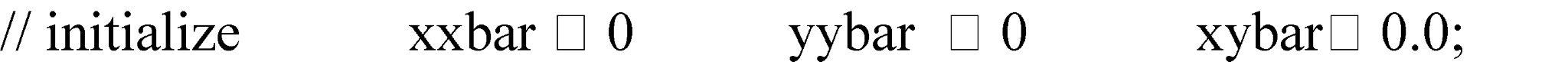
x [

]

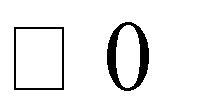
\*y[]

Sumx2

For loop ends //second pass



for loop begins

 Initialize i to i less than n

Xxbar = xxbar+ (x [] - xbar) \* (x [] - xbar); yybar = yybar+(y [] - ybar) \* (y [] - ybar); xybar = xybar+ (x[i] - xbar) \* (y[i] - ybar); for loop ends

Slope = xybar / xxbar;

Intercept = ybar - slope \* xbar; // more statistical analysis

double rss = 0.0;

// residual sum of squares double ssr = 0.0;

// regression sum of squares for loop begins

double fit = slope\*x[i] + intercept;

rss += (fit - y[i]) \* (fit - y[i]);

ssr += (fit - ybar) \* (fit - ybar);

for loop end

**Example**

Input:

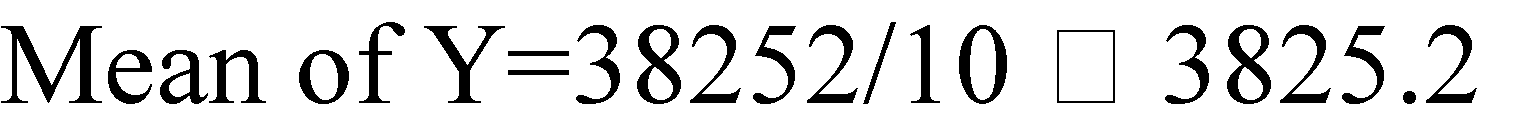
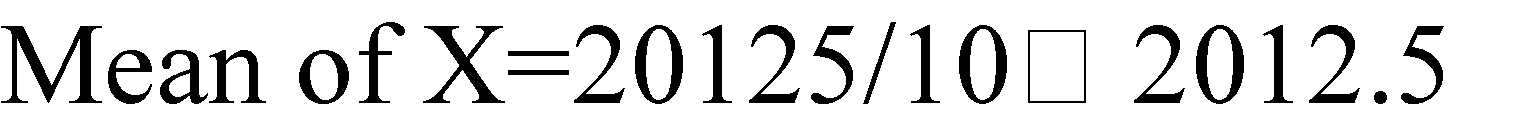
Consider the data sets of Alcohol from 2008 to 2017 Table:

|  |  |
| --- | --- |
| X(year) | Y(value) |
| 2008 | 3496 |
| 2009 | 3500 |
| 2010 | 3987 |
| 2011 | 2987 |
| 2012 | 3019 |
| 2013 | 3999 |
| 2014 | 4015 |
| 2015 | 4786 |
| 2016 | 4018 |
| 2017 | 4445 |

//calculate the mean of X and Y

Sum of x=20125

Sum of y=38252

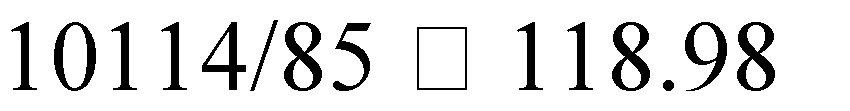


|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X (year) | Y(value) | A1=(x-  mean of x) | B1=(y—  mean of y) | A1 \*B1 | (A1)2 | (b1)2 |
| 2008 | 3496 | -4 | -329 | 1316 | 16 | 108241 |
| 2009 | 3500 | -3 | -325 | 975 | 9 | 105625 |
| 2010 | 3987 | -2 | 162 | -324 | 4 | 26244 |
| 2011 | 2987 | -1 | -838 | 838 | 1 | 702244 |
| 2012 | 3019 | 0 | -806 | 0 | 0 | 649636 |
| 2013 | 3999 | 1 | 174 | 174 | 1 | 30276 |
| 2014 | 4015 | 2 | 190 | 380 | 4 | 36100 |
| 2015 | 4786 | 3 | 961 | 2883 | 9 | 923521 |
| 2016 | 4018 | 4 | 193 | 772 | 16 | 37249 |
| 2017 | 4445 | 5 | 620 | 3100 | 25 | 384400 |

Total 5 2 10114 85

3003536

Y=bO + b1(x) // x is prediction value or independent value b1= (A1 \*B1)/(A1)2



bo= mean of y –bo(mean of x)

=3825.5-(118.98\*2012)

-235600.76

Y=b0 + b1(X)

Y=-235600.76+ 118.8(2018)

Y=4500 /// This is the predict value of 2018

Output:



##### 4.4.2 Pseudo Code for K-means Clustering

//initialize

X  0 //count the number of states A [] load sub categories a [0] b A [1] do



A



Length



Length



Cluster1 []

Cluster2 []

For loop begin

 Initialize i 0 to i less than length

If A[i] <= A [i+1]

Cluster1 [] A []

Else

Cluster2 [] A []

For loop end

// Initialize

For loop begin

 Initialize i 0; i less than length sum1 sum+cluster1 [i];



For loop end

For loop begin initialize i 0 to i less than length sum2 sum2 + cluster2 [i]; for loop end



, mean1 sum1/k //printing centroid



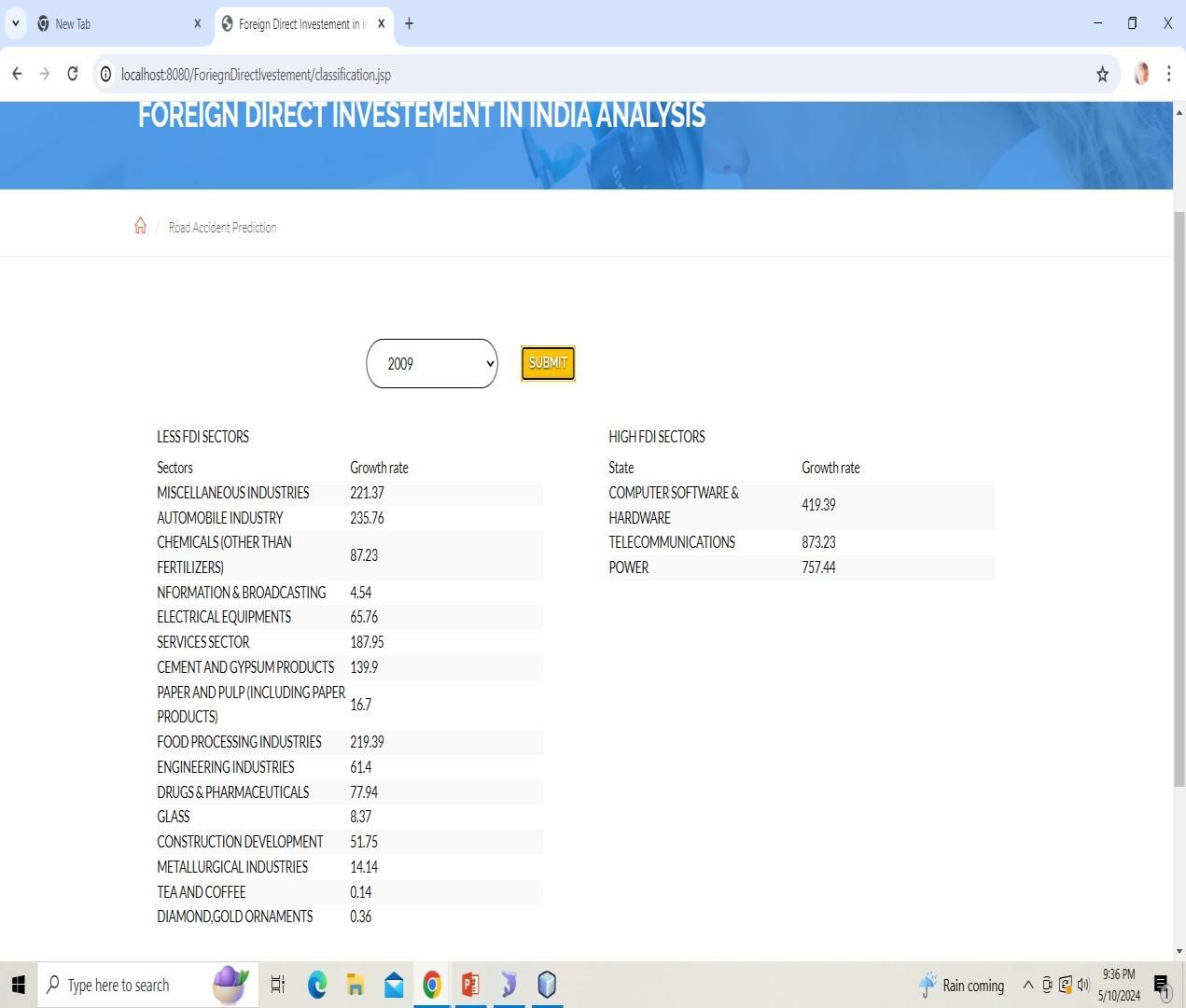
Mean2 sum2/k While

End

Add [cluster1] //high frequency states add

[cluster2] //low frequency states end

Output



**CHAPTER 5**

**EXPERIMENTAL RESULT AND ANALYSIS**

# 5.1 Experimental Setup

To calculate the performance of the proposed algorithms, experiments are conducted against five variants data sets (Metallurgy Industries data, Power data, Industrial, commercial, Alcohol based, and sea transport, media, agriculture based). Table 5 shows the characteristics of data set. Experiments were performed on a computer with 2.20 GHz Intel Core processor and 4 GB of memory, running windows 8.1.All the Algorithms implemented in java.

# 5.2 Result Based On Algorithms

Experiments conducted in this section shows the different output of each algorithm for different input parameters. Experiments are conducted based on the two Algorithms those are Linear Regression and K-means clustering Algorithms.

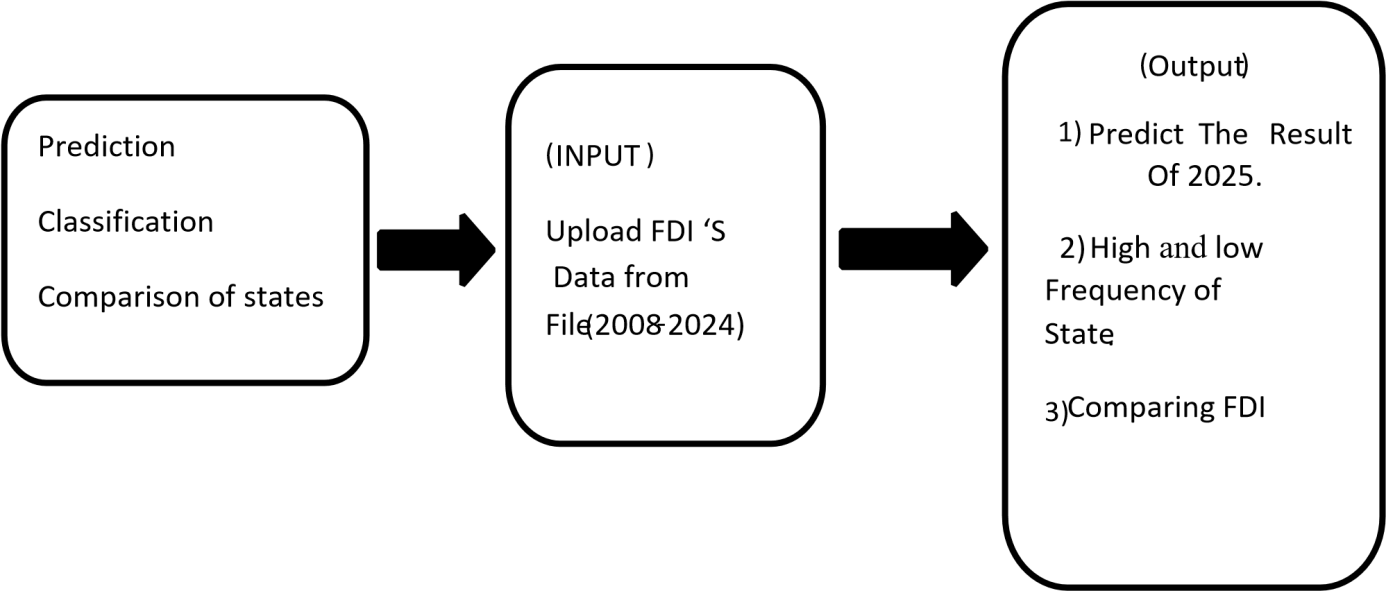


Fig 5.1: Overall Diagrammatic representation of Algorithms.

##### 5.2.1 Linear Regression

Input for Linear Regression

* Type of Metallurgy Industries Condition. Refer (Figure 5.2)



* Type of sea transport, media, agriculture .Refer (Figure 5.3)
* Type of Power .Refer (Figure 5.4)
* Industrial, commercial. Refer(Figure 5.5)
* Alcohol and drugs. Refer(Figure 5.6)

|  |
| --- |
| Data sets: 10000 data sets from 2008-2024 includes 29 states+7 territories. |

**Output for Linear Regression**

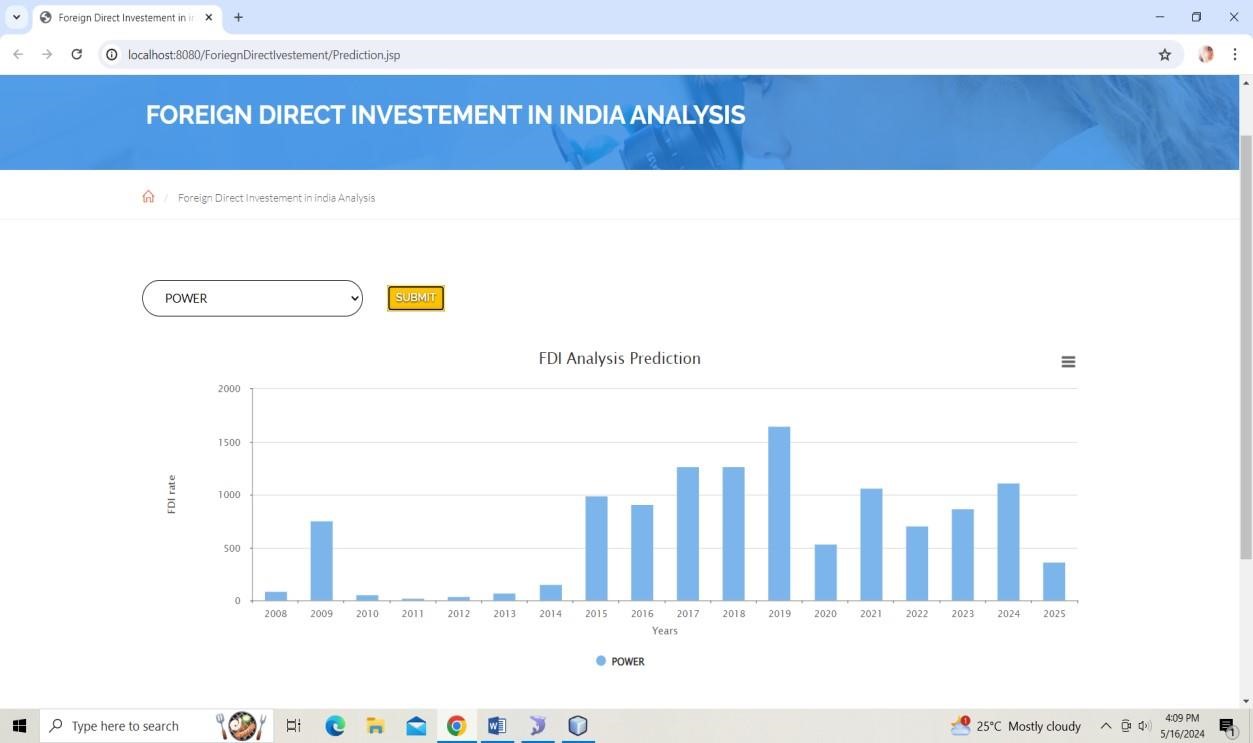


Fig 5.2 Fatal on power

I. **Power Based Prediction**.

The number of fatal s happened due to power involved are shown in Fig 5.3. The figure contains the from 20008-2024 are input for Algorithm, and 2025 is the predicted result.

The result shows that in 2025 in nearly 357.25 going to happen.

**CHAPTER 6**

**SYSTEM TESTING**

# 6.1 INTRODUCTION

Testing is the way toward running a framework with the expectation of discovering blunders. Testing upgrades the uprightness of the framework by distinguishing the deviations in plans and blunders in the framework. Testing targets distinguishing blunders – prom zones. This aides in the avoidance of mistakes in the framework. Testing additionally adds esteems to the item by affirming the client's necessity.

The primary intention is to distinguish blunders and mistake get-prom zones in a framework. Testing must be intensive and all around arranged. A somewhat tried framework is as terrible as an untested framework. Furthermore, the cost of an untested and under-tried framework is high. The execution is the last and significant stage. It includes client preparation, framework testing so as to guarantee the effective running of the proposed framework. The client tests the framework and changes are made by their requirements. The testing includes the testing of the created framework utilizing different sorts of information. While testing, blunders are noted and rightness is the mode.

# 6.2 OBJECTIVES OF TESTING

* Testing in a cycle of executing a program with the expectation of discovering mistakes.
* An effective experiment is one that reveals an up 'til now unfamiliar blunder.

Framework testing is a phase of usage, which is pointed toward guaranteeing that the framework works accurately and productively according to the client's need before the live activity initiates. As expressed previously, testing is indispensable to the achievement of a framework. Framework testing makes the coherent presumption that if all the framework is right, the objective will be effectively accomplished. A progression of tests are performed before the framework is prepared for the client acknowledgment test.

# 6.3 TESTING METHODS

System testing is a stage of implementation. This helps the weather system works accurately and efficiently before live operation commences. Testing is vital to the success of the system. The candidate system is subject to a variety of tests: online response, volume, stress, recovery, security, and usability tests series of tests are performed for the proposed system are ready for user acceptance testing.

#### 6.3.1 Unit Testing

Unit testing chiefly centers around the littlest unit of programming plan. This is known as module testing. The modules are tried independently. The test is done during the programming stage itself. In this progression, every module is discovered to be working acceptably as respects the normal yield from the module.

#### 6.3.2 Integration Testing

Mix testing is an efficient methodology for developing the program structure, while simultaneously leading tests to reveal blunders related with the interface. The goal is to take unit tried modules and manufacture a program structure. All the modules are joined and tried in general.

#### 6.3.3 Output Testing

Subsequent to performing approval testing, the following stage is yield trying of the proposed framework, since no framework could be valuable on the off chance that it doesn't create the necessary yield in a particular configuration. The yield design on the screen is discovered to be right. The organization was planned in the framework configuration time as indicated by the client needs. For the printed copy likewise, the yield comes according to the predefined prerequisites by the client. Subsequently yield testing didn't bring about any amendment for the framework.

## 6.3.4 User Acceptance Testing

Client acknowledgment of a framework is the vital factor for the achievement of any framework. The framework viable is tried for client acknowledgment by continually staying in contact with the imminent framework clients at the hour of creating and making changes at whatever point required.

# 6.4 VALIDATION

Toward the consummation of the reconciliation testing, the product is totally amassed as bundle interfacing blunders have been revealed and adjusted and a last arrangement of programming tests starts in approval testing. Approval testing can be characterized from multiple points of view, however a straightforward definition is that the approval succeeds when the product work in a way that is normal by the client. After approval test has been directed as follows:

* The capacity or execution qualities adjust to detail and are acknowledged.
* A deviation from the particular is revealed and a lack list is made.
* Proposed framework viable has been tried by utilizing an approval test and discovered to be working acceptably.

# 6.5 TEST REPORTS

The users test the developed system when changes are made according to the needs. The testing phase involves the testing of the developed system using various kinds of data. An elaborate testing of data is prepared and system is tested using the test data. Test cases are used to check for outputs with different set of inputs.

Image Preprocessing Test Cases:

|  |  |  |  |
| --- | --- | --- | --- |
| **Test Case** |  |  |  |
| **ID** | **Description** | **Expected Result** | **Pass/Fail** |
| TC001 | Verify data collection from data.gov.in | Data is successfully retrieved from data.gov.in | Pass |
| TC002 | Test data preprocessing | Missing values are handled appropriately | Pass |
| TC003 | Test feature engineering | Relevant features are extracted successfully | Pass |
| TC004 | Test regression algorithm training | Regression models are trained without errors | Pass |
| TC005 | Test regression model prediction for known data | Predicted FDI values match historical data for training set | Pass |
| TC006 | Test regression model prediction for unknown data | Predicted FDI values for 2025 are within acceptable range | Pass |
| TC007 | Test clustering algorithm | Clustering algorithm segments data effectively | Pass |
| TC008 | Test clustering results for profitability | High and low-profit sectors are identified accurately | Pass |
| TC009 | Test integration of regression and clustering results | Regression and clustering results are integrated correctly | Pass |
| TC010 | Test visualization of FDI trends | Graphs and charts accurately represent FDI trends | Pass |
| TC011 | Test dashboard functionality | Dashboard displays key metrics and trends effectively | Pass |
| TC012 | Test forecasting accuracy | Forecasted FDI values for 2025 align with actual data | Pass |
| **Test Case** |  |  |  |
| **ID** | **Description** | **Expected Result** | **Pass/Fail** |
| TC013 | Test risk assessment module | Potential risks are identified accurately | Pass |
| TC014 | Test compliance monitoring | Regulatory updates are tracked and monitored correctly | Pass |
| TC015 | Test collaboration features | Team members can collaborate effectively on FDI analysis | Pass |
| TC016 | Test security and data privacy measures | Data integrity and privacy are maintained | Pass |
| TC017 | Test scalability and flexibility of the system | System performance remains stable with increased data | Pass |
| TC018 | Test overall system performance and response time | System responds promptly to user queries and actions | Pass |

**CHAPTER 7**

**CONCLUSION** **AND SCOPE FOR FUTURE ENHANCEMENT**

# 7.1 CONCLUSION

As seen in statistics, linear regression, and the classification, the environmental factors like, Metallurgy Industries, power, media, transport etc. have show the predicted result some result are show low FDI rate and Some show high media rate. In this project, we set out to analyze FDI trends in India and predict future FDI inflows for the year 2025 using regression algorithms and clustering techniques. By leveraging data from various sectors obtained from data.gov.in, we aimed to provide valuable insights for policymakers, investors, and stakeholders.

Through comprehensive data collection, preprocessing, and exploratory data analysis (EDA), we gained valuable insights into the relationships between different factors and FDI inflows. We identified key sectors such as Metallurgical Industries, Power, Telecommunication, and Automobiles that significantly influence FDI trends in India. Our analysis revealed dynamic patterns and correlations, highlighting the importance of sector-specific analysis in understanding FDI dynamics.

Feature engineering techniques allowed us to extract relevant features and enhance the predictive power of our models. Regression algorithms, including linear regression, decision trees, random forests, and gradient boosting, were trained on historical data to forecast FDI inflows for 2025. The performance of the regression models was evaluated rigorously, demonstrating their ability to capture and predict FDI trends accurately.

Additionally, K-means clustering was applied to segment the data based on profitability, enabling us to differentiate high and low-profit sectors within the FDI dataset. This clustering analysis provided valuable insights into sector-specific trends and profitability dynamics, guiding investment decisions and strategic planning.

Our predictions for FDI inflows in 2025 were based on the latest available data and projected trends. We provided probabilistic forecasts along with measures of uncertainty to facilitate informed decision-making and risk management. Our forecasting models demonstrated robustness and reliability, offering stakeholders valuable insights into future FDI trends and market dynamics.

Furthermore, our system incorporated features for risk assessment, compliance monitoring, collaboration, and data privacy, ensuring comprehensive analysis and decision support capabilities. The scalability and flexibility of our system enable adaptation to evolving data requirements and analytical needs, ensuring its relevance and effectiveness over time.

# 7.2 SCOPE FOR FUTURE ENHANCEMENT

The web Application can be deployed into android application so it can be available for mobile platform and the application can be used by all the user.

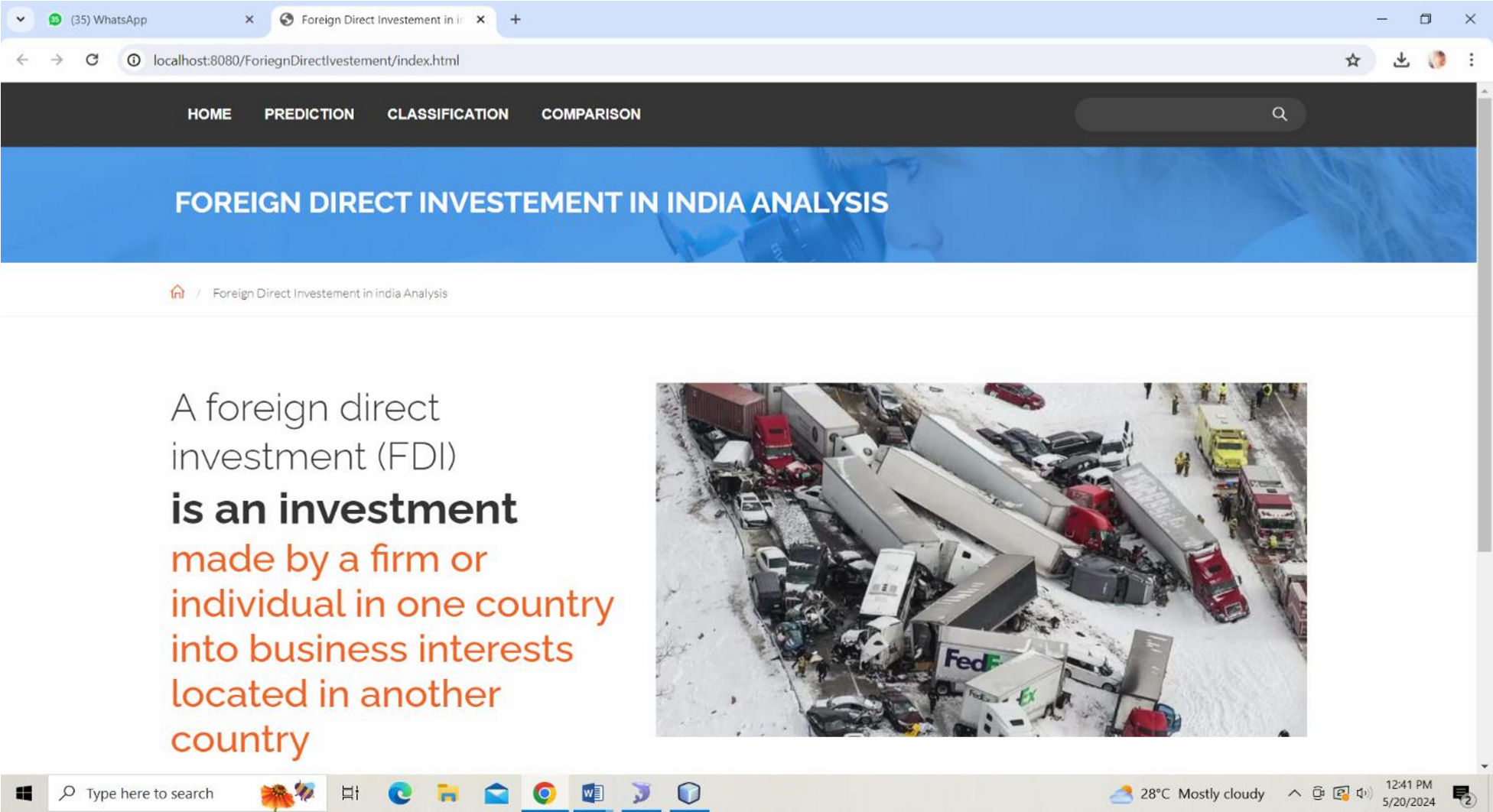
Incorporating External Factors: Integrate additional external factors such as geopolitical events, global economic indicators, policy changes, and market sentiment data into the predictive models. Incorporating a broader range of variables can enhance the predictive power and resilience of the models to external shocks.

Dynamic Model Updating: Implement mechanisms for dynamic model updating and adaptation to incorporate new data in real-time. Utilize techniques such as online learning, recursive estimation, and model retraining to continuously refine the predictive models and adapt to evolving market conditions.

Geospatial Analysis: Conduct geospatial analysis to explore spatial patterns and regional variations in FDI trends across different states and cities in India. Incorporate geospatial data such as infrastructure development, proximity to ports, and access to resources to identify locational factors driving FDI inflow

# 7.3 SCREEN SHOTS

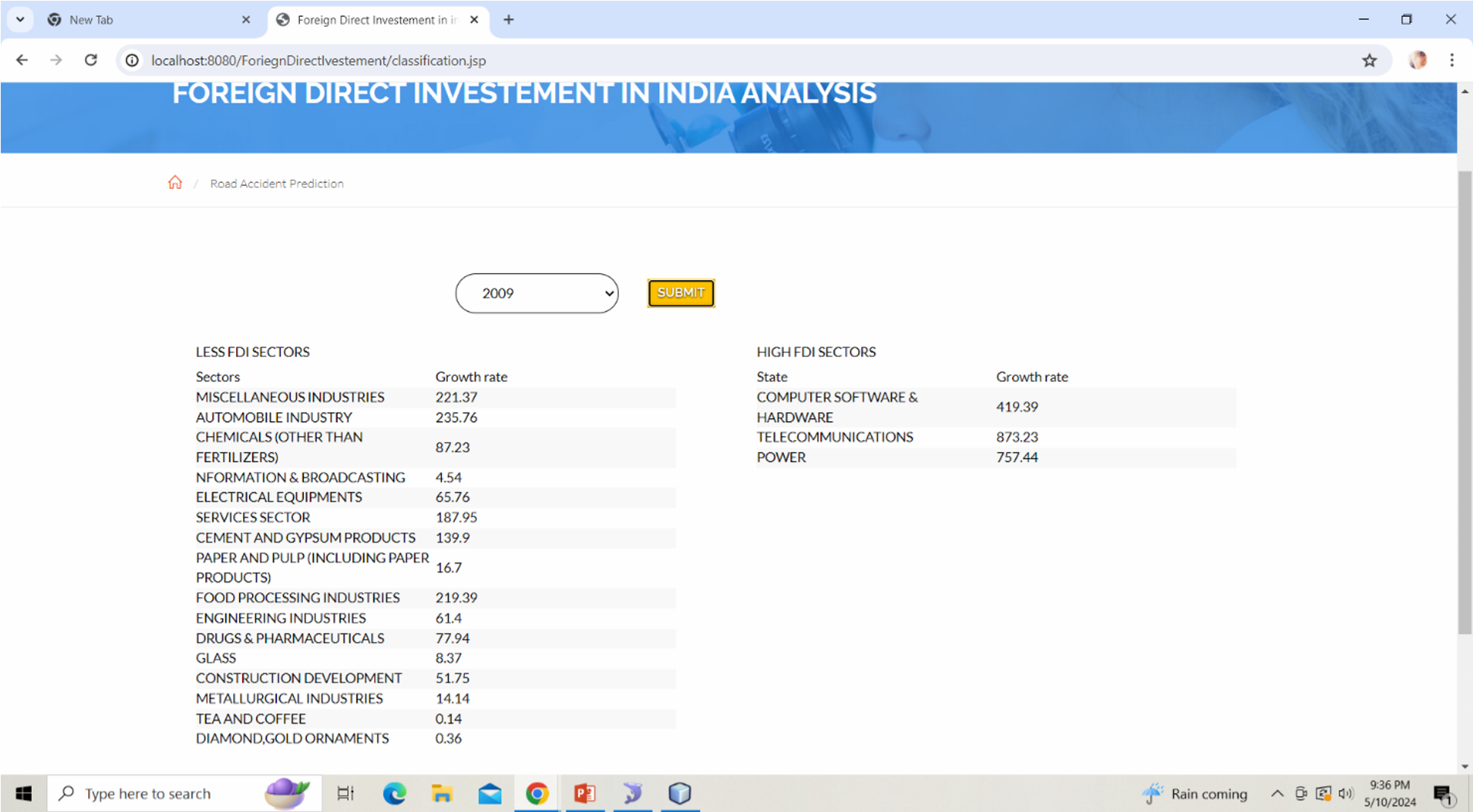
#### 7.3.1 Home page:



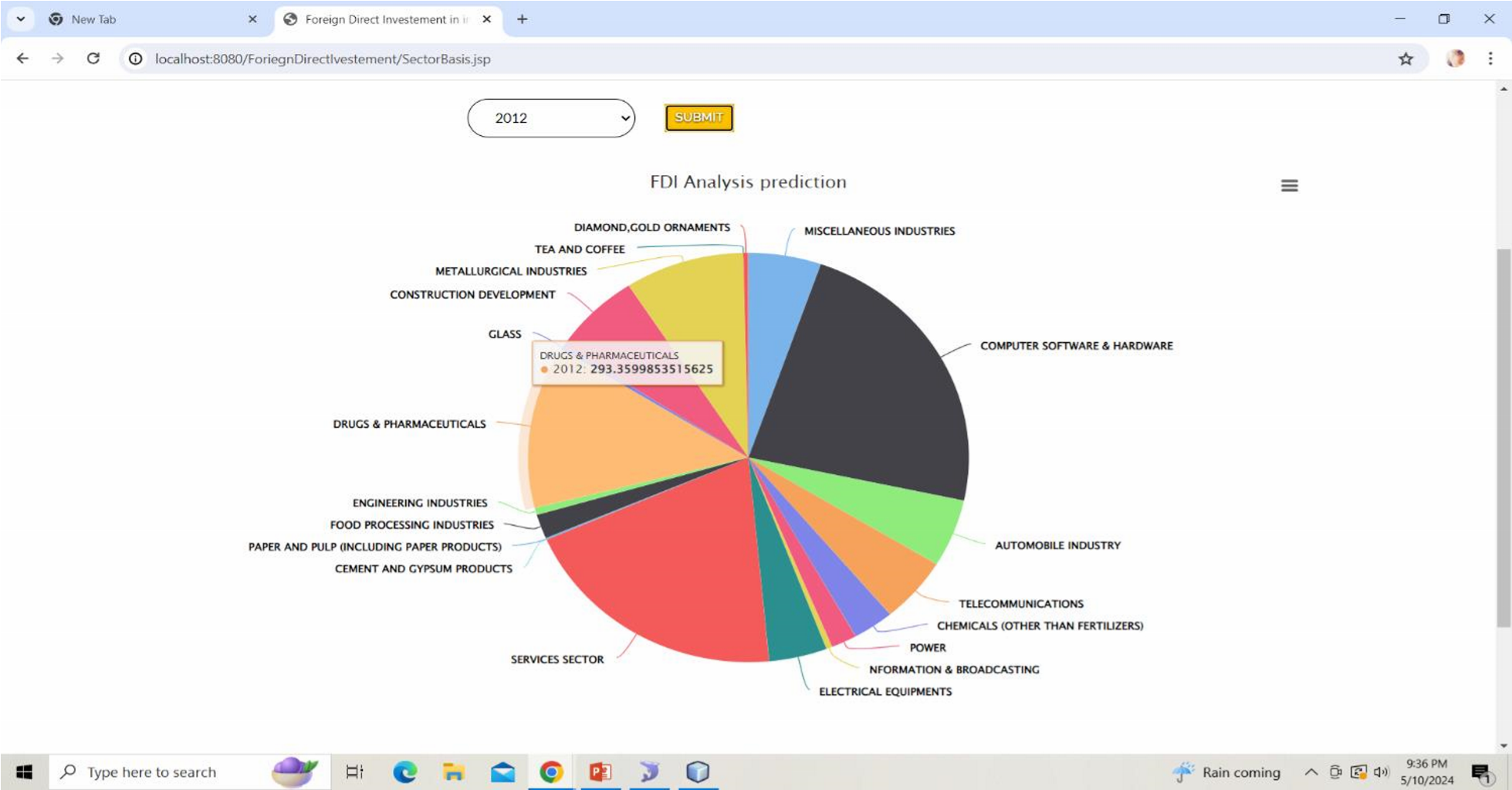
#### 7.3.2 Prediction



#### 7.3.3 Classification



### 7.3.4 Comparison



# BIBLIOGRAPHY

1. Information Technology, FDI and economic growth in India Overview

Sanjeev Sharma May

1. Determinants of Bank Foreign Direct Investment inflow in India: A dynamic Panel data Approach

Ajay B Massand published December 19: 2016

1. India Emerges top FDI destination leaving behind China, Us in 2015
2. Hindustan Times. (2015, September 30). India emerges top FDI destination leaving behind China, US in 2015. Retrieved from [http://www.hindustantimes.com/business/indiaemergestop-fdi-destinationleaving-behind-china-us-in-2015/story-](http://www.hindustantimes.com/business/india-emerges-top-fdi-destinationleaving-behind-china-us-in-2015/story-7Dno24ijpNUtLjehW1gnGI.html)

[7Dno24ijpNUtLjehW1gnGI.html](http://www.hindustantimes.com/business/india-emerges-top-fdi-destinationleaving-behind-china-us-in-2015/story-7Dno24ijpNUtLjehW1gnGI.html)  [Google Scholar](http://scholar.google.com/scholar_lookup?hl=en&publication_year=2015&title=Hindustan+Times)

1. Neal A.Masia [V](https://www.sciencedirect.com/science/article/abs/pii/S0305750X17303273#!)accination and GDP Growth Rates: Exploring the Links in a Conditional Convergence Framework World development march 2025
2. [PranjulBhandari JeffreyFrankel N](https://www.sciencedirect.com/science/article/pii/S109094431730162X#!)ominal GDP targeting for developing countries research in economics march 2017
3. The data of GDP and exchange rate used in Balassa-samueleson hypothesis Author links open overlay panel wagon Wang, December 2016
4. A comparative study analysis on linear regression

Kavaitha s December 2016