ADBMS_UNIT-5

Introduction to Data Mining

Data mining is the process of extracting useful information from large sets of data. It involves using various techniques from statistics, machine learning, and database systems to identify patterns, relationships, and trends in the data. This information can then be used to make data-driven decisions, solve business problems, and uncover hidden insights. Applications of data mining include customer profiling and segmentation, market basket analysis, anomaly detection, and predictive modeling. Data mining tools and technologies are widely used in various industries, including finance, healthcare, retail, and telecommunications.

Applications of Data Mining

1. Financial Analysis
2. Biological Analysis
3. Scientific Analysis
4. Intrusion Detection
5. Fraud Detection
6. Research Analysis

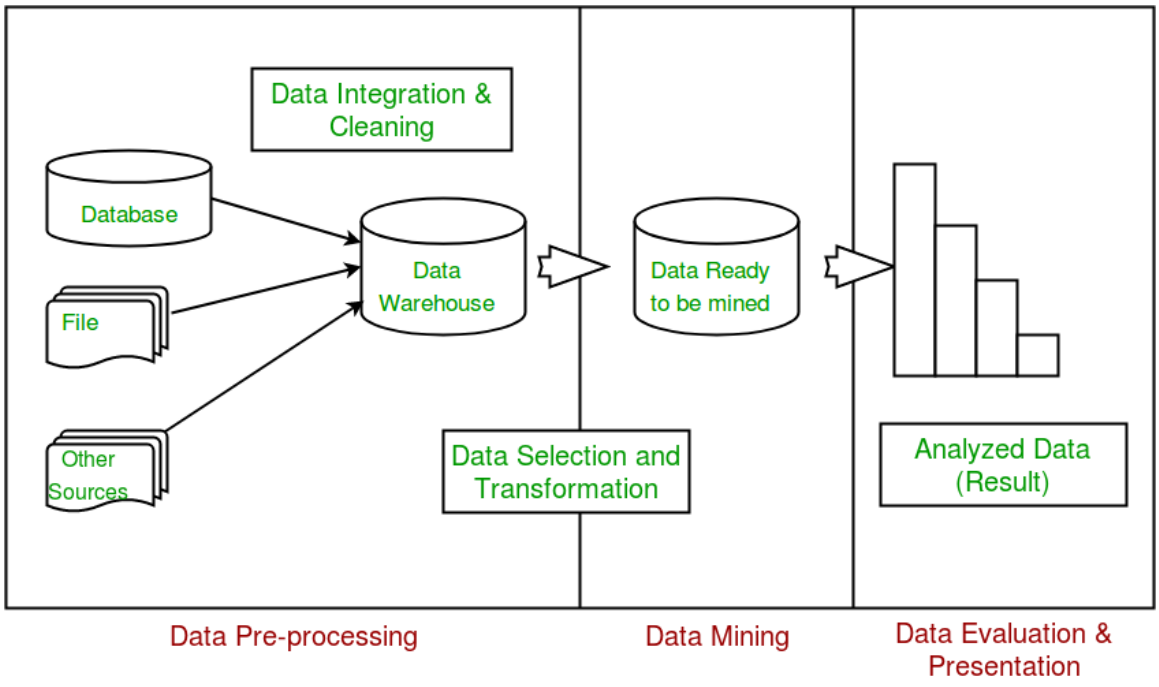
Benefits of Data Mining

1. Improved decision-making by identifying patterns and trends in large data sets.
2. Enhanced competitiveness through uncovering new business opportunities.
3. Fraud detection by identifying anomalies and unusual patterns.
4. Predictive modeling to forecast future events and trends.
5. Risk management by analyzing customer behavior and market conditions.

Data Mining as a Whole Process

The whole process of Data Mining consists of three main phases:

1. Data Pre-processing – Data cleaning, integration, selection, and transformation takes place
2. Data Extraction – Occurrence of exact data mining
3. Data Evaluation and Presentation – Analyzing and presenting results



KDD Process / Data Mining Implementation

KDD (Knowledge Discovery in Databases) is a process that involves the extraction of useful, previously unknown, and potentially valuable information from large datasets. The KDD process is an iterative process and it requires multiple iterations of the above steps to extract accurate knowledge from the data. The following steps are included in KDD process:

Data Cleaning

Data cleaning is defined as removal of noisy and irrelevant data from collection.

1. Cleaning in case of **Missing values**.

2. Cleaning **noisy** data, where noise is a random or variance error.
3. Cleaning with **Data discrepancy detection** and **Data transformation tools**.

Data Integration

Data integration is defined as heterogeneous data from multiple sources combined in a common source(Data Warehouse). Data integration using **Data Migration tools, Data Synchronization tools and ETL**(Extract-Load-Transformation) process.

Data Selection

Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection. For this we can use **Neural network, Decision Trees, Naive bayes, Clustering**, and **Regression** methods.

Data Transformation

Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure. Data Transformation is a two step process:

1. **Data Mapping**: Assigning elements from source base to destination to capture transformations.
2. **Code generation**: Creation of the actual transformation program.

Data Mining

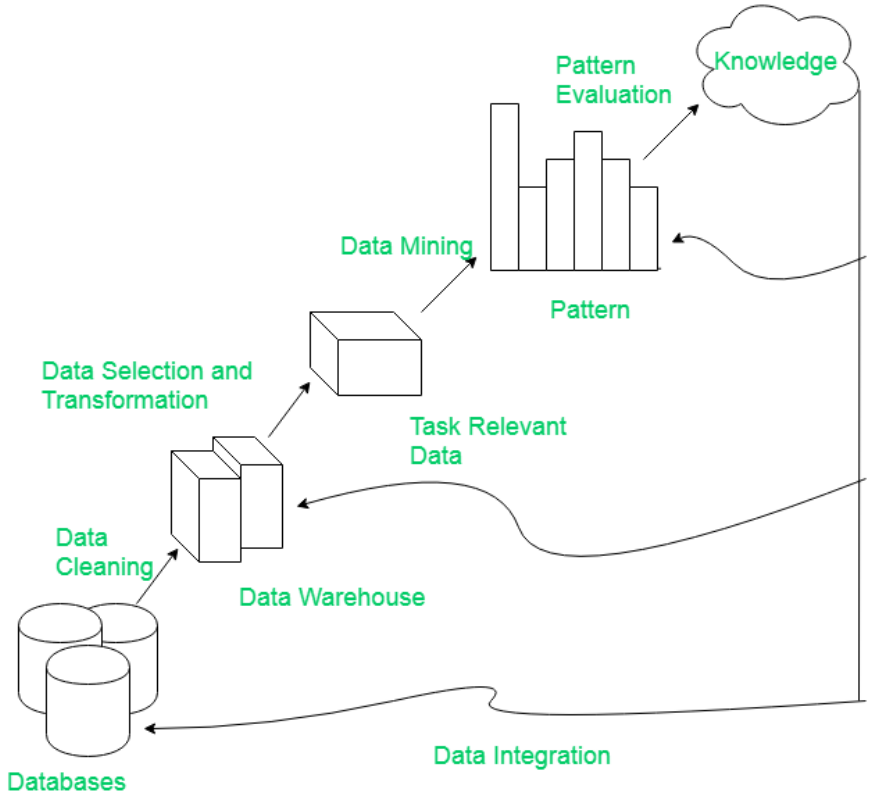
Data mining is defined as techniques that are applied to extract patterns potentially useful. It transforms task relevant data into **patterns, and d**ecides purpose of model using **classification** or **characterization**.

Pattern Evaluation

Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures. It find **interestingness score** of each pattern, and uses **summarization** and **Visualization** to make data understandable by user.

Knowledge Representation

This involves presenting the results in a way that is meaningful and can be used to make decisions.



Advantages of KDD

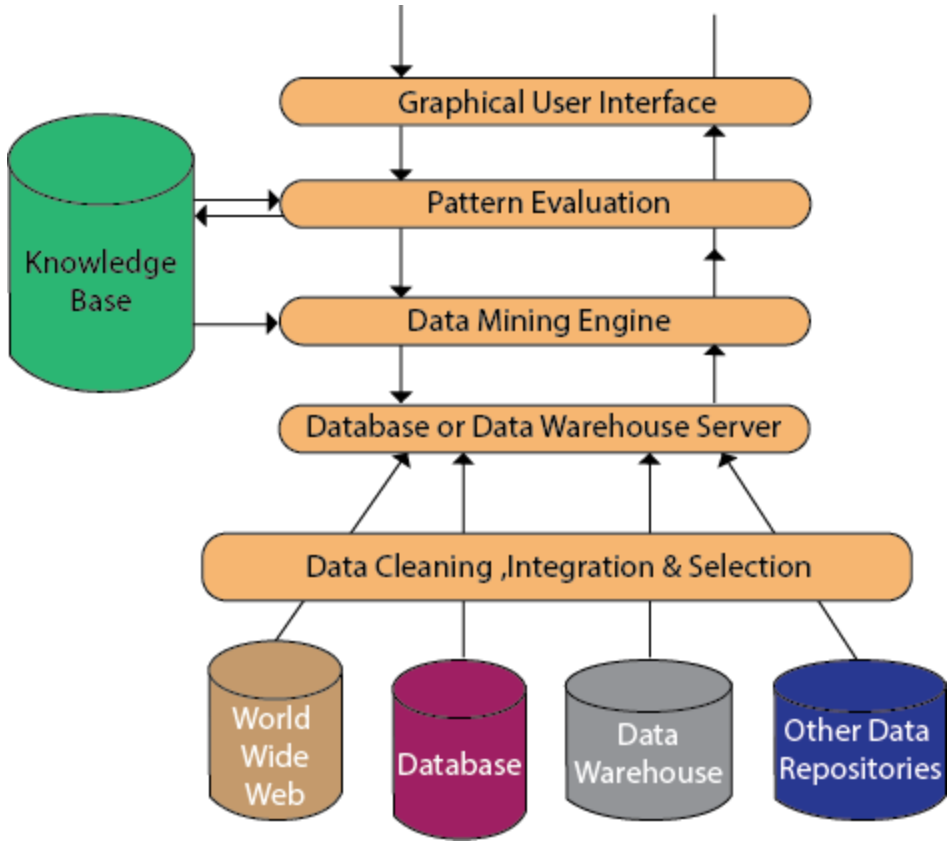
1. **Improves decision-making:** KDD provides valuable insights and knowledge that can help organizations make better decisions.
2. **Increased efficiency:** KDD automates repetitive and time-consuming tasks and makes the data ready for analysis, which saves time and money.
3. **Better customer service:** KDD helps organizations gain a better understanding of their customers' needs and preferences, which can help them provide better customer service.
4. **Fraud detection:** KDD can be used to detect fraudulent activities by identifying patterns and anomalies in the data that may indicate fraud.
5. **Predictive modeling:** KDD can be used to build predictive models that can forecast future trends and patterns.

Disadvantages of KDD

1. **Privacy concerns:** KDD can raise privacy concerns as it involves collecting and analyzing large amounts of data, which can include sensitive information about individuals.
2. **Complexity:** KDD can be a complex process that requires specialized skills and knowledge to implement and interpret the results.
3. **Unintended consequences:** KDD can lead to unintended consequences, such as bias or discrimination, if the data or models are not properly understood or used.
4. **Data Quality:** KDD process heavily depends on the quality of data, if data is not accurate or consistent, the results can be misleading
5. **High cost:** KDD can be an expensive process, requiring significant investments in hardware, software, and personnel.
6. **Overfitting:** KDD process can lead to overfitting, which is a common problem in machine learning where a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new unseen data.

Data Mining Architecture

The significant components of data mining systems are a data source, data mining engine, data warehouse server, the pattern evaluation module, graphical user interface, and knowledge base.



Data Source:

The actual source of data is the Database, data warehouse, World Wide Web (WWW), text files, and other documents. You need a huge amount of historical data for data mining to be successful. Organizations typically store data in databases or data warehouses. Data warehouses may comprise one or more databases, text files spreadsheets, or other repositories of data. Sometimes, even plain text files or spreadsheets may contain information. Another primary source of data is the World Wide Web or the internet.

Different processes:

Before passing the data to the database or data warehouse server, the data must be cleaned, integrated, and selected. As the information comes from various sources and in different formats, it can't be used directly for the data mining procedure because the data may not be complete and accurate. So, the first data requires to be cleaned and unified. More information than needed will be collected from various data sources, and only the data of interest will have to be selected and passed to the server. These procedures are not as easy as we think. Several

methods may be performed on the data as part of selection, integration, and cleaning.

Database or Data Warehouse Server:

The database or data warehouse server consists of the original data that is ready to be processed. Hence, the server is cause for retrieving the relevant data that is based on data mining as per user request.

Data Mining Engine:

The data mining engine is a major component of any data mining system. It contains several modules for operating data mining tasks, including association, characterization, classification, clustering, prediction, time-series analysis, etc.

In other words, we can say data mining is the root of our data mining architecture. It comprises instruments and software used to obtain insights and knowledge from data collected from various data sources and stored within the data warehouse.

Pattern Evaluation Module:

The Pattern evaluation module is primarily responsible for the measure of investigation of the pattern by using a threshold value. It collaborates with the data mining engine to focus the search on exciting patterns.

This segment commonly employs stake measures that cooperate with the data mining modules to focus the search towards fascinating patterns. It might utilize a stake threshold to filter out discovered patterns. On the other hand, the pattern evaluation module might be coordinated with the mining module, depending on the implementation of the data mining techniques used. For efficient data mining, it is abnormally suggested to push the evaluation of pattern stake as much as possible into the mining procedure to confine the search to only fascinating patterns.

Graphical User Interface:

The graphical user interface (GUI) module communicates between the data mining system and the user. This module helps the user to easily and efficiently use the system without knowing the complexity of the process. This module

cooperates with the data mining system when the user specifies a query or a task and displays the results.

Knowledge Base:

The knowledge base is helpful in the entire process of data mining. It might be helpful to guide the search or evaluate the stake of the result patterns. The knowledge base may even contain user views and data from user experiences that might be helpful in the data mining process. The data mining engine may receive inputs from the knowledge base to make the result more accurate and reliable. The pattern assessment module regularly interacts with the knowledge base to get inputs, and also update it.

Difference Between Descriptive and Predictive Data Mining

Descriptive mining:

This term is basically used to produce correlation, cross-tabulation, frequency etc. These technologies are used to determine the similarities in the data and to find existing patterns. One more application of descriptive analysis is to develop the captivating subgroups in the major part of the data available. This analytics emphasis on the summarization and transformation of the data into meaningful information for reporting and monitoring.

Examples of descriptive data mining include clustering, association rule mining, and anomaly detection. Clustering involves grouping similar objects together, while association rule mining involves identifying relationships between different items in a dataset. Anomaly detection involves identifying unusual patterns or outliers in the data.

Predictive Data Mining:

The main goal of this mining is to say something about future results not of current behaviour. It uses the supervised learning functions which are used to predict the target value. The methods come under this type of mining category are called classification, time-series analysis and regression. Modelling of data is

the necessity of the predictive analysis, and it works by utilizing a few variables of the present to predict the future not known data values for other variables.

Examples of predictive data mining include regression analysis, decision trees, and neural networks. Regression analysis involves predicting a continuous outcome variable based on one or more predictor variables. Decision trees involve building a tree-like model to make predictions based on a set of rules. Neural networks involve building a model based on the structure of the human brain to make predictions.

The main differences between descriptive and predictive data mining are:

Purpose: Descriptive data mining is used to describe the data and identify patterns and relationships. Predictive data mining is used to make predictions about future events.

Approach: Descriptive data mining involves analyzing historical data to identify patterns and relationships. Predictive data mining involves using statistical models and machine learning algorithms to identify patterns and relationships that can be used to make predictions.

Output: Descriptive data mining produces summaries and visualizations of the data. Predictive data mining produces models that can be used to make predictions.

Timeframe: Descriptive data mining is focused on analyzing historical data. Predictive data mining is focused on making predictions about future events.

Applications: Descriptive data mining is used in applications such as market segmentation, customer profiling, and product recommendation. Predictive data mining is used in applications such as fraud detection, risk assessment, and demand forecasting.

Difference Between Descriptive and Predictive Data Mining:

S.No.	Comparison	Descriptive Data Mining	Predictive Data Mining
1.	Basic	It determines, what happened in the past	It determines, what can happen in the future with

		by analyzing stored data.	**the help past data analysis.**
2.	**Preciseness**	**It provides accurate data.**	**It produces results does not ensure accuracy.**
3.	**Practical analysis methods**	**Standard reporting, query/drill down and ad-hoc reporting.**	**Predictive modelling, forecasting, simulation and alerts.**
4.	**Require**	**It requires data aggregation and data mining**	**It requires statistics and forecasting methods**
5.	**Type of approach**	**Reactive approach**	**Proactive approach**
6.	**Describe**	**Describes the characteristics of the data in a target data set.**	**Carry out the induction over the current and past data so that predictions can be made.**
7.	**Methods(in general)**	**• what happened? • where exactly is the problem? • what is the frequency of the problem?**	**• what will happen next? • what is the outcome if these trends continue? • what actions are required to be taken?**

Data Mining Tools

1. Orange Data Mining:

Orange is a perfect machine learning and data mining software suite. It supports the visualization and is a software-based on components written in Python computing language and developed at the bioinformatics laboratory at the faculty of computer and information science, Ljubljana University, Slovenia.

As it is a software-based on components, the components of Orange are called "widgets." These widgets range from preprocessing and data visualization to the assessment of algorithms and predictive modeling.

Widgets deliver significant functionalities such as:

- Displaying data table and allowing to select features

- Data reading
- Training predictors and comparison of learning algorithms
- Data element visualization, etc.

Besides, Orange provides a more interactive and enjoyable atmosphere to dull analytical tools. It is quite exciting to operate.

Why Orange?

Orange is an open-source data visualization and analysis tool that supports both beginners and professionals. It offers a user-friendly interface with visual programming (drag-and-drop widgets) and Python scripting for data mining tasks. The platform supports over 100 widgets for operations such as reading data, feature selection, training predictors, and visualizing results with tools like bar charts, scatterplots, and heat maps. It is compatible with Windows, Mac OS X, and Linux, and includes machine learning components, bioinformatics add-ons, and text mining features for comprehensive data analysis.

Orange supports multiple regression and classification algorithms, focusing on supervised data mining techniques. It allows for the creation of ensembles by combining predictions from individual models, enhancing precision. Ensembles and other models can be diversified by altering parameter sets or using different training data. Python integration enables seamless operation in environments like PyCharm and iPython. Orange's flexibility and extensive features make it a powerful tool for making smarter decisions by rapidly comparing and analyzing data.

2. SAS Data Mining

- **What It Does**: SAS (Statistical Analysis System) is a powerful tool for data mining, data management, and analytics. It can process and analyze data from multiple sources, transform data, and generate statistical insights.
- **Why It Is Used**: SAS is ideal for handling big data and producing actionable insights for timely decision-making. It supports data mining, text mining, and optimization tasks with high scalability. Its distributed memory processing architecture makes it suitable for large-scale data operations.
- **Tech Stack**:

- Base: Proprietary software.
- Language: SAS scripting language.
- Integration: Supports various databases and sources for data integration.

---

3. DataMelt (DMelt)

- **What It Does**: DMelt is a computation and visualization environment designed for interactive data analysis and visualization. It includes libraries for scientific and mathematical operations like 2D/3D plotting, curve fitting, and random number generation.
- **Why It Is Used**: Popular among scientists, engineers, and students for analyzing large datasets and performing statistical and computational analysis. It is extensively used in natural sciences, financial markets, and engineering.
- **Tech Stack**:
  - Base: Java-based application compatible with JVM.
  - Libraries: Scientific libraries (for plots) and mathematical libraries (for algorithms, fitting, etc.).
  - Platform: Multi-platform (Windows, Mac, Linux).

---

4. Rattle

- **What It Does**: Rattle provides a GUI for data mining tasks and uses the R programming language for statistical analysis. It allows users to process data, visualize results, and generate code for further customization and extension.
- **Why It Is Used**: Rattle simplifies the use of R's statistical power for non-programmers through its GUI, making it suitable for beginners and experts alike. It supports editing datasets, reviewing generated code, and extending functionality for advanced tasks.
- **Tech Stack**:
  - Base: R programming language.
  - Libraries: Leverages R's statistical and visualization libraries.

- Interface: GUI-based with integrated code-generation capabilities.

5. Rapid Miner

- **What It Does**: Rapid Miner is a comprehensive platform for predictive analytics, machine learning, deep learning, and text mining. It provides an integrated environment for various data analysis and mining tasks.
- **Why It Is Used**: Ideal for enterprise applications, research, and education due to its template-based framework for rapid deployment with fewer errors. Its client/server model supports on-site, public, and private cloud deployment.
- **Tech Stack**:
  - Base: Java-based application.
  - Libraries: Built-in libraries for machine learning and deep learning.
  - Deployment: Supports both on-site and cloud infrastructures.

Data Mining Applications

Scientific Analysis: Scientific simulations are generating bulks of data every day. This includes data collected from nuclear laboratories, data about human psychology, etc. Data mining techniques are capable of the analysis of these data. Now we can capture and store more new data faster than we can analyze the old data already accumulated. Example of scientific analysis:

- Sequence analysis in bioinformatics
- Classification of astronomical objects
- Medical decision support.

Intrusion Detection: A network intrusion refers to any unauthorized activity on a digital network. Network intrusions often involve stealing valuable network resources. Data mining technique plays a vital role in searching intrusion detection, network attacks, and anomalies. These techniques help in selecting and refining useful and relevant information from large data sets. Data mining technique helps in classify relevant data for Intrusion Detection System. Intrusion Detection system generates alarms for the network traffic about the foreign invasions in the system. For example:

- Detect security violations
- Misuse Detection
- Anomaly Detection

Business Transactions: Every business industry is memorized for perpetuity. Such transactions are usually time-related and can be inter-business deals or intra-business operations. The effective and in-time use of the data in a reasonable time frame for competitive decision-making is definitely the most important problem to solve for businesses that struggle to survive in a highly competitive world. Data mining helps to analyze these business transactions and identify marketing approaches and decision-making. Example :

- Direct mail targeting
- Stock trading
- Customer segmentation
- Churn prediction (Churn prediction is one of the most popular Big Data use cases in business)

Market Basket Analysis: Market Basket Analysis is a technique that gives the careful study of purchases done by a customer in a supermarket. This concept identifies the pattern of frequent purchase items by customers. This analysis can help to promote deals, offers, sale by the companies and data mining techniques helps to achieve this analysis task. Example:

- Data mining concepts are in use for Sales and marketing to provide better customer service, to improve cross-selling opportunities, to increase direct mail response rates.
- Customer Retention in the form of pattern identification and prediction of likely defections is possible by Data mining.
- Risk Assessment and Fraud area also use the data-mining concept for identifying inappropriate or unusual behavior etc.

Education: For analyzing the education sector, data mining uses Educational Data Mining (EDM) method. This method generates patterns that can be used both by learners and educators. By using data mining EDM we can perform some educational task:

- Predicting students admission in higher education
- Predicting students profiling
- Predicting student performance
- Teachers teaching performance
- Curriculum development
- Predicting student placement opportunities

Research: A data mining technique can perform predictions, **classification**, clustering, associations, and grouping of data with perfection in the research area. Rules generated by data mining are unique to find results. In most of the technical research in data mining, we create a training model and testing model. The training/testing model is a strategy to measure the precision of the proposed model. It is called Train/Test because we split the data set into two sets: a training data set and a testing data set. A training data set used to design the training model whereas testing data set is used in the testing model. Example:

- Classification of uncertain data.
- Information-based clustering.
- Decision support system
- Web Mining
- Domain-driven data mining
- IoT (Internet of Things)and Cybersecurity
- Smart farming IoT(Internet of Things)

Healthcare and Insurance: A Pharmaceutical sector can examine its new deals force activity and their outcomes to improve the focusing of high-value physicians and figure out which promoting activities will have the best effect in the following upcoming months, Whereas the Insurance sector, data mining can help to predict which customers will buy new policies, identify behavior patterns of risky customers and identify fraudulent behavior of customers.

- Claims analysis i.e which medical procedures are claimed together.
- Identify successful medical therapies for different illnesses.

- Characterizes patient behavior to predict office visits.

Transportation: A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services. A large consumer merchandise organization can apply information mining to improve its business cycle to retailers.

- Determine the distribution schedules among outlets.
- Analyze loading patterns.

Financial/Banking Sector: A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product.

- Credit card fraud detection.
- Identify ‘Loyal’ customers.
- Extraction of information related to customers.
- Determine credit card spending by customer groups.