**Project Title: Decision Support Portal (a case study of Loan department at International Bank Liberia Limited)**

A project report submitted in partial fulfilment of the degree of

**B.Sc. Information Technology**

**Submitted by:**

*Abraham M. Sheriff*

**Name of Supervisor:**

*Mr. Shyam Nair*

**Abraham M. Sheriff**

**302921610160**



**DEPARTMENT OF INFORMATION TECHNOLOGY**

**BLUECREST UNIVERSITY COLLEGE, LIBERIA**

**(Affiliated by National Commision on Higher Education, Liberia)**

**Fall 2021**

**Dedication**

To my Late Mother

And friend,

Near and Far

**Certificate**

This is to certify that this project report entitled “**Decision Support Portal**” *by***Abraham M. Sheriff (302921610160),**submitted in partial fulfillment of the requirements for the degree of B.Sc. Information Technology in the BlueCrest University College, Liberia, during the academic year 2020,is a bonafide record of work carried out under my guidance and supervision.

**Supervisor Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Approved By**

**Head of Academics: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Place: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Declaration**

I, the undersigned, solemnly declare that the project report DECISION SUPPORT PORTAL is based on my own work carried out during the course of my study under the supervision of Mr. Shyam Nair.

I assert the statements made and conclusions drawn are an outcome of my research work. I further certify that:

1. The work contained in the report is original and has been done by me under the supervisor.
2. The work has not been submitted to any other Institution for any other degree/diploma/certificate in any other university in Liberia or abroad.
3. I have followed the guidelines provided by the university college in writing the report.
4. Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have cited them in the text of the report and provided their details in the references.

**Acknowledgement**

No tasks is single man’s effort. It is done through cooperation and coordination of other people. Based on that, I would like to acknowledge the contribution of the following personalities in guiding me for bringing up this project successfully.

* Mr. Shyam Nair –head of Department for Information Technology (IT ), BlueCrest University College Liberia

Prince Kennedy Mark Oguche, Faculty BlueCrest University College Liberia.

List of Table

1. Introduction
2. Literature Review
3. Research Methodology
4. System Analysis & Design
5. Recommendation and Conclusion

# List of Figures

[Chapter Contents](#_Toc525726633)

Cover Page ….….….….……………………………………...…………………..i  
      Certificate….……………………………………………………………………..ii

Declaration….……………………………………………………………………iii

Internship certificate…….….….….….….….….….….……………………………………..iv  
        Acknowledgement….…………………………………………………………… v

List of Tables….…………………………………………………………………vi  
       List of Figures…………………………………………………………………..vii

Abstract………………………………………………………………………….viii

[1.0 Introduction….….….….….….….….….….….….….….….….….….….….….….…1](#_Toc525726634)

[1.1. Background of the Study](#_Toc525726634) -------------------------------------------------------------------------1-2

1.2. Research Problem Statement -------------------------------------------------------------------3

1.3. Significance of Study ---------------------------------------------------------------------- --------4-5

1.4 Research Objective & Hypothesis ---------------------------------------------------------------6-8

1.5 Research Questions ---------------------------------------------------------------------------------9

1.6 Delimitation of Study -------------------------------------------------------------------------------10

[2. Literature Review (Tabular Format) ……………..…………………………….……](#_Toc525726635)11-24

3.0 Research Methodology……………………………………………..…………….25

3.1 Introduction-------------------------------------------------------------------------------------------25-26

3.2 Design method ---------------------------------------------------------------------------------------27

3.3 Justification of method -----------------------------------------------------------------------------28

3.4 Tools and techniques --------------------------------------------------------------------------------29

[4.0 Data Analysis and System Design………………………………………………………](#_Toc525726637)30

4.1 Introduction --------------------------------------------------------------------------------------------31

4.2 Overview of Proposed System ---------------------------------------------------------------------32

4.3 System Design ---------------------------------------------------------------------------------------32-34

4.3.1 Feasibility Study------------------------------------------------------------------------35

4.3.1.1 Technical Feasibility------------------------------------------------------36

4.3.1.2 Organizational Feasibility-----------------------------------------------37

4.3.1.3 Economic Feasibility------------------------------------------------------38

4.4 Requirement Planning------------------------------------------------------------------------------39

4.4.1 Functional Requirement--------------------------------------------------------------40

4.4.2 Non-functional Requirement--------------------------------------------------------41

4..4.3 System Requirement-----------------------------------------------------------------42

4.5 System Implementation (Screenshots or layout of Project)-----------------------------43-58

4.6 System Testing---------------------------------------------------------------------------------------59

[5.0 Conclusions and Recommendation….…………………………………………….…](#_Toc525726638)60

5.1 Conclusion----------------------------------------------------------------------------------------------62-64

5.2 Key Contribution--------------------------------------------------------------------------------------63

5.3 Limitations & Future Works -----------------------------------------------------------------------65-67

5.4 References----------------------------------------------------------------------------------------------68

5.5 Appendices---------------------------------------------------------------------------------------------69

**Abstract**

The core concept of Decision Support Portal is centering on the organization of data in charts or tabular form in order to track performance of the business by means of information. These raw data keep operational officers aware of what is happening with the business. When the business is not reaching one of its goals, the reporting charts should alert you of the issue, prompting you to respond. And the analysis of these data, is a process of inspecting, cleansing, transforming and modeling data with the aim objective of discovering useful information, inform conclusion and supporting decision-making. So with that, the decision support portal will be using some predefined queries for generating reports that will help business intelligence or financial analyst in decision-making processes, particularly in classifying customer for loan application in banks or any other financial institutions.

**Introduction**

Decision support portal is a data mining and data warehousing toolkit that help management or loan officers in classifying and categorizing loan’s applicant in banks or some other mortgage related entities as either “Risky” or “Safe”. with rapid growth of data in today’s industrial world, there is a need for development of databases and informational portals for easy retrieval of data that will enhance intelligence decision-making in financial institutions, telecommunication, insurance, health etc. this is applicable in a scenario where there is an availability of huge amount of data from various operational data sources like flat-files and transactional databases for drawing inferences. With the help of these data sources, data can be stored in huge or bulk in difference kinds of databases or information repositories, which will help in retrieving large volume of data that will enable end-users to carry on sophisticated analysis, reporting and predicting future trends in any given scenario.

More interestingly, research has shown that a considerable number of entities in Liberia do not have a data mart nor a data Warehouse for mining or extracting knowledge for hidden, valid and potentially useful pattern in huge data set.

* 1. **Background of the study**

Indeed, the concept of data reporting and analysis has gain a tremendous admiration by institutions for generating report. Which help analyst for intelligence decision making.

The exponentially increasing amounts of data being generated each year make getting useful information from that data more and more critical. The information frequently is stored in a data warehouse, a repository of data gathered from various sources, including corporate databases, summarized information from internal systems, and data from external sources. Analysis of the data includes simple query and reporting, statistical analysis, more complex multidimensional analysis, and data mining.

Data analysis and reporting are a subset of business intelligence (BI), which also incorporates data warehousing, database management systems, and Online Analytical Processing (OLAP).

The technologies are frequently used in customer relationship management (CRM) to analyze patterns and query customer databases. Large quantities of data are searched and analyzed to discover useful patterns or relationships, which are then used to predict future behavior.

Some estimates indicate that the amount of new information doubles every three years. To deal with the mountains of data, the information is stored in a repository of data gathered from various sources, including corporate databases, summarized information from internal systems, and data from external sources. Properly designed and implemented, and regularly updated, these repositories, called data warehouses, allow managers at all levels to extract and examine information about their company, such as its products, operations, and customers' buying habits.

With a central repository to keep the massive amounts of data, organizations need tools that can help them extract the most useful information from the data. A data warehouse can bring together data in a single format, supplemented by metadata through use of a set of input mechanisms known as extraction, transformation, and loading (ETL) tools. These and other Business Intelligence tools enable organizations to quickly make knowledgeable business decisions based on good information analysis from the data.

Analysis of the data includes simple query and reporting functions, statistical analysis, more complex multidimensional analysis, and data mining (also known as knowledge discovery in databases, or KDD). Online analytical processing (OLAP) is most often associated with multidimensional analysis, which requires powerful data manipulation and computational capabilities.

With the increasing data being produced each year, Data Reporting has become a hot topic. The increasing focus on business intelligence has caused a number of large organizations have begun to increase their presence in the space, leading to a consolidation around some of the largest software vendors in the world. Among the notable purchases in the business intelligence market were Oracle's purchase of Hyperion Solutions; Open Text's acquisition of Hummingbird.

* 1. **Research Problem Statements**

While there has been a tremendous change in data reporting and analysis by means of data mining and data warehousing in tech hub nation like Americas, Europe, and some part of Africa. Third world countries like Liberia, stay applying the traditional approach of data reporting and analysis for business intelligences. This is the clear indication that there is a need for data warehouse in some of these areas for easy reporting.

Furthermore, to implement such conventional, mature and trusted discipline task for specific business unit or data warehouse, there are many pending issues that have to be addressed. There are several implementation issues and challenges that need to be tangle.

The like of data sources – my main goal in my data collection was to have some preliminary data on hand. With the research would have been more reliable, accurate and balance like quantitative data.

* 1. **Significance of the Study**

Primarily, The purpose of gathering corporate information together in a single structure, typically an organization's data warehouse, is to facilitate analysis so that information that has been collected from a variety of different business activities may be used to enhance the understanding of underlying trends in their business. Analysis of these data can include simple query and reporting functions, statistical analysis, more complex multidimensional analysis, and data mining. OLAP, one of the fastest growing areas, is most often associated with multidimensional analysis. According to The business Verdict formerly The OLAP Report, the definition of the characteristics of an OLAP application is "fast analysis of shared multidimensional information.

Data warehouses are usually separated from production systems, as the production data is added to the data warehouse at intervals that vary, according to business needs and system constraints. Raw production data must be cleaned and qualified, so it often differs from the operational data from which it was extracted. The cleaning process may actually change field names and data characters in the data record to make the revised record compatible with the warehouse data rule set. This is the province of Extract, Transform and Loan (ETL).

A data warehouse also contains metadata (structure and sources of the raw data, essentially, data about data), the data model, rules for data aggregation, replication, distribution and exception handling, and any other information necessary to map the data warehouse, its inputs, and its outputs. As the complexity of data analysis grows, so does the amount of data being stored and analyzed; ever more powerful and faster analysis tools and hardware platforms are required to maintain the data warehouse.

A successful data warehousing strategy requires a powerful, fast, and easy way to develop useful information from raw data. Data analysis and data mining tools use quantitative analysis, cluster analysis, pattern recognition, correlation discovery, and associations to analyze data with little or no Information Technology (IT) intervention. The resulting information is then presented to the user in an understandable form, processes collectively known as Business Intelligence. Managers can choose between several types of analysis tools, including queries and reports, managed query environments, and OLAP and its variants (ROLAP, MOLAP, and HOLAP). These are supported by data mining, which develops patterns that may be used for later analysis, and completes the decision making process.

Moreover, with the increasing number of large data in institutions, organizations or entities, storage and easy retrieval of some these data from large set of data has been computational expensive. It does require the involvement or participation of skilled computer engineer most especially a Data Scientist for extraction of knowledge from some of these data set.

In addition to that, in order for data to really be valuable to an organization, we need to be able to discover patterns and relationships within that data. And that’s what data mining does. Those connections and insights can enable better business decisions. Data mining can also reduce risk, helping us to detect fraud, errors, and inconsistencies that can lead to profit loss and reputation damage. Different industries use data mining in different contexts, but the goal is the same: to better understand customers and the business. Nowadays, data mining is widely use in area like E-ecommerce, education, Telecommunication, banks, healthcare for generating report and intelligence decision-making. The first example of Data Mining and Business Intelligence comes from service providers in the mobile phone and utilities industries. They organize billing information, customer services interactions, website visits and other metrics to give each customer a probability score, then target offers and incentives to customers whom they perceive to be at a higher risk of churning.

* 1. **Research Objectives and Hypothesis**

With the increasing need of data reporting, Data mining has been one of the most widely used methods to extract data from diverse sources and organize them for better usage. In spite of having different commercial systems for data mining, a lot of challenges come up when they are actually implemented. With rapid evolution in the field of data mining, companies are expected to stay abreast with all the new developments.

Complex algorithms form the basis for data mining as they allow for data segmentation to identify various trends and patterns, detect variations, and predict the probabilities of various events happening. The raw data may come in both analog and digital format, and is inherently based on the source of the data. Companies need to keep track of the latest data mining trends and stay updated to do well in the industry and overcome challenging competition.

Furthermore, based on some of these trends, I hypothesized the tremendous change in the following technologies.

**Multimedia Data Mining**

This is one of the latest methods which is catching up because of the growing ability to capture useful data accurately. It involves the extraction of data from different kinds of multimedia sources such as audio, text, hypertext, video, images, etc. and the data is converted into a numerical representation in different formats. This method of data mining can be used in clustering and classifications, performing similarity checks, and also to identify associations.

**Ubiquitous Data Mining**

Ubiquitous Data Mining is the process of extracting hidden classifier, cluster, frequent item-set, and association rules from distributed data among a number of mobile and stationary data sources.

This method involves the mining of data from mobile devices to get information about individuals. In spite of having several challenges in this type such as complexity, privacy, cost, etc. this method has a lot of opportunities to be enormous in various industries especially in studying human-computer interactions.

**Distributed Data Mining**

This is another future scenario in technologies, as the world continues to undergo a rapid change in technologies particularly in the area of Artificial intelligence, Machine Learning and Internet of Things.

This type of data mining is gaining popularity as it involves the mining of huge amount of information stored in different company locations or at different organizations. Highly sophisticated algorithms are used to extract data from different locations and provide proper insights and reports based upon them.

**Spatial and Geographic Data Mining**

This is another new trending type of data mining which includes extracting information from environmental, astronomical, and geographical data which also includes images taken from outer space. This type of data mining can reveal various aspects such as distance, is mainly used in geographic information systems and other navigation applications.

**Time Series and Sequence Data Mining**

The primary application of this type of data mining is study of cyclical and seasonal trends. This practice is also helpful in analyzing even random events which occur outside the normal series of events. This method is mainly being use by retail companies to access customer's buying patterns and their behaviors.

In addition to that, in order for data to really be valuable to an organization, we need to be able to discover patterns and relationships within that data. And that is what data mining does. Those connections and insights can enable better business decisions. Data mining can also reduce risk, helping us to detect fraud, errors, and inconsistencies that can lead to profit loss and reputation damage. Different industries use data mining in different contexts, but the goal is the same: to better understand customers and the business. Nowadays, data mining is widely use in area like E-ecommerce, education, Telecommunication, banks, healthcare for generating report and intelligence decision-making. The first example of Data Mining and Business Intelligence comes from service providers in the mobile phone and utilities industries. They organize billing information, customer services interactions, website visits and other metrics to give each customer a probability score, then target offers and incentives to customers whom they perceive to be at a higher risk of churning.

**Retail:**

Another example of Data Mining and Business Intelligence comes from the retail sector. Retailers segment customers into ‘Recency, Frequency, Monetary’ (RFM) groups and target marketing and promotions to those different groups. A customer who spends little but often and last did so recently will be handled differently to a customer who spent big but only once, and also some time ago. The former may receive a loyalty, upsell and cross-sell offers, whereas the latter may be offered a win-back deal, for instance.

**E-commerce:**

Perhaps some of the most well-known examples of Data Mining and Analytics come from E-commerce sites. Many E-commerce companies use Data Mining and Business Intelligence to offer cross-sells and up-sells through their websites. One of the most famous of these is, of course, Amazon, who use sophisticated mining techniques to drive their, people who viewed their product, also liked this functionality.

* 1. **Research Questions**

1. How to make Data Analysis and Reporting process automated instead of manual?
2. What is Data Analysis & Reporting?
3. Why Data mining & Data Warehousing are gaining popularity?
4. When should we implement the knowledge of data mining & data warehousing?
   1. **Delimitation of the Study**

Primarily, this research project confine itself in interviewing and observing the loan officers from International Bank Liberia Limited credit department on considerable factors for granting loan like customer’s income, collateral value, age and other miscellaneous expenses.

1. **Literature Review (Tabular Form)**

|  |  |
| --- | --- |
| **Related Works** | **Author** |
| 1. **Improve your Financial Report** | **Emil Dr.Kusic** |
| 1. **Customers Segmentation** | **Er. Kiranbir Kaur’s** |
| 1. **Data Mining Techniques in Analyzing a Process Data- A Didactic** | **Xin Qiao & Hong Jiao** |
| 1. **Market Basket Analysis Using Data Mining** | **Malisa Santos** |
| 1. **Data Mining on Customer Segmentation** | **Er. Rupampreet Kaur & Er. Kiranbir Kaur** |

**Emil DrKusic’s research papers, “financial reporting and analysis”**

In Emil DrKusic’s research papers, “financial reporting and analysis” <https://www.vertabelo.com/blog/improve-your-financial-reporting-with-data-warehousing/> , he tries discussing on how to build a data mart as a solution to a difficult data reporting and aggregation. Also, he tries transforming data and migrate it from external systems to a data warehouse, which was basically meant for storing bulk of information or data from diverse sources.

One of the benefits of his technique that I do concord with him is that, his data warehouse structure will remain the same regardless of changes to operational applications or regulations. This way, anyone who has to generate reports will be able to use predefined queries or report templates. In that, they will spend less time in dealing with the technicalities of report generation and have more time to use their expertise to interpret results. On the contrary, the users may have little or no working knowledge of Structured Query Language (SQL), the native language of Relational Database Management System (RDBMS) for querying data. And in Emil DrKusic’s DataMart for financial reporting, he made the end-users to use their intuition in interpreting their results, which I think is computational expensive for just mere end-user. So in that, I think he should have used his programming skill to design a well-defined, responsive and interactive user interface for the end-users of his financial report for the DataMart.

And as developer, I would like to achieve one of the eight golden rules of interface design – to reduce burdensome on user’s memory. By designing responsive, interactive and learnable user-interface for the end user of the data mart.

**Customers’ segmentation using R**

This concept of data mining is the process of dividing customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing research such as gender, age, interests, and miscellaneous spending habits.

In this Data Science R Project series, the developers perform one of the most essential applications of machine learning – Customer Segmentation. In this project, he implemented customer segmentation using R. Whenever you need to find your best customer, customer segmentation is the ideal methodology.

In this machine learning project, Data Flair will provide you the background of customer segmentation. Then the user will explore the data upon which the will be building segmentation model. Also, in this data science project, they saw the descriptive analysis of data and then implement several versions of the K-means algorithm.

Customer Segmentation is one the most important applications of unsupervised learning. Using clustering techniques, companies can identify the several segments of customers allowing them to target the potential user base. In this machine learning project, the developer make use of K-means Clustering algorithm. Which is the essential algorithm for clustering unlabeled dataset.

Companies that deploy customer segmentation are under the notion that every customer has different requirements and require a specific marketing effort to address them appropriately. Companies aim to gain a deeper approach of the customer they are targeting. Therefore, their aim has to be specific and should be tailored to address the requirements of each and every individual customer. Furthermore, through the data collected, companies can gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that would reap them maximum profit. This way, they can strategize their marketing techniques more efficiently and minimize the possibility of risk to their investment.

The technique of customer segmentation is dependent on several key differentiators that divide customers into groups to be targeted. Data related to demographics, geography, economic status as well as behavioral patterns play a crucial role in determining the company direction towards addressing the various segments.

**Xin Qiao & Hong Jiao’s work on Data Mining**

Well, Xin Qiao & Hong Jiao’s article, There study analyzed the process data in the log file from one of the 2012 PISA problem-solving items using data mining techniques. The data mining methods used, including CART, gradient boosting, random forest, SVM, SOM, and *k*-means, yielded satisfactory results with this dataset. The three major purposes of the current study are summed in the following words.

First, to demonstrate the analysis of process data using both supervised and unsupervised techniques, concrete steps in feature generation, feature selection, classifier development and outcome evaluation were presented in the current study. Among all steps, feature generation was the most crucial one because the quality of features determines the classification results to a large extent. Good features should be created based on a thorough understanding of the item scoring procedure and the construct. Key action sequences that can distinguish correct and incorrect answers served as features with good performance. Unexpectedly, time features, including total response time and its pieces, did not turn out to be important features for classification. This means that considerable variance of response time existed in each score group and the differences in response time distributions among the groups was not large enough to clearly distinguish the groups. This study generated features based on theoretical beliefs about the construct measured and used students as the unit of analysis. The data could be structured in other ways according to different research questions. For example, instead of using students as the unit of analysis, the attempts students made can be used as rows and actions as columns, then the attempts can be classified instead of people.

Second, to evaluate classification consistency of these frequently used data mining techniques, the current study compared four supervised techniques with different properties, namely, CART, gradient boosting, random forest, and SVM. All four methods achieved satisfactory classification accuracy based on various outcome measures, with gradient boosting showing slightly better overall accuracy and Kappa value. In general, easy interpretability and graphical visualization are the major advantages of trees. Trees also deal with noisy and incomplete data well. However, the trees are easily influenced by even small changes in the data due to its hierarchical splitting structure. SVM, on the contrary, generalizes well because once the hyper-plane is found, small changes to data cannot greatly affect the hyper-plane. Given the specific dataset in the current study, even the CART method worked very well. In addition, the CART method can be easily understood and provided enough information about the detailed classifications between and within each score category. Thus, based on the results in the current study, the CART method is sufficient for future studies on similar datasets. Unsupervised learning algorithms, SOM and *k*-means, also showed convergent clustering results based on DBI and Kappa values. In the final clustering solution, students were grouped into 9 clusters, revealing specific problem-solving processes they went through.

Third, supervised and unsupervised learning methods serve to answer different research questions. Supervised learning methods can be used to train the algorithm to predict memberships in the future data, like automatic scoring. Unsupervised methods can reveal the problem-solving strategy patterns and further differentiate students in the same score category. This is especially helpful for formative purposes. Students can be provided with more detailed and individualized diagnostic reports. Teachers can better understand students' strengths and weaknesses, and adjust instructions in the classroom accordingly or provide more targeted tutoring to specific students. In addition, it is necessary to check any indication for cheating behavior in the misclassified or outlier cases from both types of data mining methods. For example, students answered the item correctly within an extremely short amount of time can imply item compromise.

This study has its own limitations. Other data mining methods, such as other decision trees algorithms and clustering algorithms, are worth of investigation. However, the procedure demonstrated in this study can be easily generalized to other algorithms. In addition, the six methods were compared based on the same set of data rather than data under various conditions. Therefore, the generalization of the current study is limited due to factors such as sample size and number of features. Future studies can use a larger sample size and extract more features from more complicated assessment scenarios. Lastly, the current study focuses on only one item for the didactic purpose. In the future study, process data for more items can be analyzed simultaneously to get a comprehensive picture of the students.

To sum up, the selection of data mining techniques for the analysis of process data in assessment depends on the purpose of the analysis and the data structure. Supervised and unsupervised techniques essentially serve different purposes for data mining with the former as a confirmatory approach while the latter as an exploratory approach.

**Market Basket Analysis using Data Mining – Melisa Santos**

Data mining is described as the extraction of hidden helpful information from a collection of huge databases, data mining is also a technique that encompasses an enormous form of applied mathematics and computational techniques like link analysis, clustering, classification, summarizing knowledge, regression analysis and so on. Data mining tools predict future trends and behaviors, permitting businesses to create knowledge-driven selections. The machine-driven, prospective analyses offered by data mining move on the far side the analyses of past events. Data mining tools provides answer to business questions that were time consuming. They search databases for hidden patterns, finding useful information that is beyond the reach of specialists.

Data mining techniques is enforced speedily on existing package and hardware platforms to reinforce the worth of existing information resources, and might be integrated with new product and systems as they’re brought. Once enforced on high performance client/server or multiprocessing computers, data mining tools will analyze huge databases to provide answers to questions such as, “What goods consumers tend to buy the most and goods that go along side with it”?

Broadly data mining can be defined as set of mechanisms and techniques, realized in software, to extract hidden information from data. However, the word hidden in this definition is important;

By the early 1990s data mining was commonly recognized as a sub process within a larger process called Knowledge Discovery in Databases or KDD ,the most commonly used definition of KDD is that of Fayyad et al as “the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data.’’ (Fayyad et al. 1996).

As such data mining should be viewed as the sub-process, within the overall KDD process, concerned with the discovery of hidden information”. Other sub-processes that form part of the KDD process are data preparation (warehousing, data cleaning, pre-processing, and so on) and the analysis of a results. For many practical purposes KDD and data mining are seen as synonymous, but technically one is a sub-process of the other. The data that data mining techniques were originally directed at was tabular data and, given the processing power available at the time, computational efficiency was of significant concern. As the amount of processing power generally available increased, processing became less of a concern and was replaced with a desire for accuracy and a desire to mine ever larger data collections. Today, in

The context of tabular data, we have a well-established range of data mining techniques available.

It is well within the capabilities of many commercial enterprises and researchers to mine in tabular form.

There is huge amount of data available in Information Industry. This data is of no use until converted into useful information. Analyzing this huge amount of data and extracting useful information from it is necessary. The extraction of information is not the only process we need to perform; it also involves other processes such as Data pre-processing( Data Cleaning, Data Integration, Data Transformation) Data Mining, Pattern Evaluation and Data Presentation. Once all these processes are over, we are now position to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, Science Exploration etc.

Through in depth research and observations carried on supermarket Malisa Santos discovered that retailers are willing to know what product is purchased with the other or if a particular products are purchased together as a group of items. Which can help in their decision making with respect to placement of a product, determining the timing and extent of promotions on product and also have a better understanding of customer purchasing habits by grouping customers with their transactions.

This project is aimed at designing and implementing a well-structured market basket analysis software tool to solve the problem stated above and compare the result to that of an existing software called WEKA.

The aim of the study was to maximize profit for the retailers by providing better services to the consumers

The objective of this study are:

Cross-Market Analysis – Data Mining performs Association/correlations between product sales.

Identifying Customer Requirements – helps in identifying the best products for different customers. It uses prediction to find the factors that may attract new customers.

Customer Profiling – helps to determine what kind of people buy what kind of products.

Data Pre-Processing

Due to the fact that the data we are getting is a raw data, raw data in the real world may be incomplete it has to be pre-processed the raw data has to go through data cleaning, data integration, data normalization, data reduction because without a quality data there will be no quality mining results.

Data cleaning: This has to do with filling of missing values, resolving of inconsistencies in the raw data.

Data integration: combining data from multiple sources and generating the user with unified view of the data.

Normalization: normalization is used to minimize or to reduce redundancy.

Data reduction: reduction of the data set that is much smaller in volume but yet yields the same analytical results

Well, the scope of the study focuses on Babcock Ventures supermarket and the scope of the project were centering on the following points:

They aim was to develop their own market basket analysis software, which can be used in Babcock University.

The software exhibit a colorful GUI (graphical user interface).

The software will be based on Apriori algorithm.

They intend to conduct a research into the various branches of science that this software will be based on, such as artificial intelligence.

They develop a software that will eventually stand out among other data mining software.

The limitations of this software will include:

Data restrictions: this is a major factor that stands in the way of the execution of this project. Since there is no data on households and individual consumers, they neglect such purchases.

Time constraints: this is also a major factor due to the fact that it can’t work on a small amount of raw data because it tends to mislead the retailer in a nut shell this software will work on large volumes of data.

**Data Mining on Customer Segmentation: the review Analysis**

In Er. Rupampreet Kaur & Er. Kiranbir Kaur’s paper on “Data Mining on Customer Segmentation”, they provided a clear definition of data mining as the process of extracting important information from the bulk of data to save and summarize it in an effective and efficient manner. That is, hidden information can be extracted from the large set of data. Their work highlighted the comparison of various technique based used in image segmentation and big data such as supervised learning – classification, prediction and unsupervised learning – association rule mining, clustering. These approaches are very important for highly customized industries that use large set of data. But still there were some other issues that was not considered the use of customer segmentation as well as the use of differential evolution is ignored in existing work to classify the data. So in near future, I recommend that, there should be use of differential evolution that which will help to protect our software.

1. **Research Methodology**
   1. Introduction

Indeed, there are a lot of research methods and technique for implementing proposed research methodologies that have been used by many researchers and scholars for carrying on or writing research papers. But as a tech-savvy, I do prefer conduction of interview for gathering reliable information from correspondent for collecting requirements or intelligences in any given entity. Based on that, there is a need for using the interview method in carry on the research methodology.

Well, in order to gain a better insight into the possibilities for categorizing loan applicant and granting loan to the deserving ones with the sole intent of revenue generation for the bank, a semi-structure interviews were conducted with eight (8) of the loan officers from International Bank Liberia Limited. And some of the loan officers and their management team was defined as user who usually grant loan at least twice a week in the bank. The interviews were used to select participants who belong to the target group within the age range of (21-55 years old). The interview was conducted in small office next to the audit department of International Bank Liberia Limited, which lasted roughly about fifteen minutes per individual. And the respond from the correspondent were recorded by note-taking, and few interviews were also filmed with consent. But on that, bulk of them preferred not to be filmed that which would have enable me to have a lot of evidence or data on hand.

Furthermore, before carrying on preliminary investigation to gather information, a well-defined questionnaire was organized. And the data were properly collected, prepared, pre-processed and summarized to get rid of all of the noisy data from the collected dataset. From there, I carried on Knowledge discovery in database (KDD) process as a technique to extract knowledge from the given set of data, where collected data were firstly cleaned and integrated, from there I did some data selection and transformation, them after that I implemented some data mining technique for identification of knowledge from the acquired dataset. Finally, I did pattern evaluation and knowledge presentation to identify interesting pattern in the mining stage.

So with that, I going to implement a decision tree algorithm in data mining in order to fully achieve this research project. In this algorithm, I will be creating a training data model that can be used for predicting the class or value of the target variable by learning sample decision rules infer from prior data – training data (behavioral of the loan applicant)

In spite of approaches used in this research. The sole purpose of this chapter is to design the methodology of the research approach through mixed types of research techniques. The approach also supports the researcher on how to come across the result findings. In this chapter, the general design of the research and method used for data collection are explained in detail. It includes three main parts.

The first topic is about research design method follow by justification of method. In addition to that, it encompasses the tools and techniques use for the development of the portal. And them, system analysis & design. Finally, feasibility study.

* 1. **Design Method**

|  |  |
| --- | --- |
| **Purpose** | This project seeks to make data reporting & Analysis easy as much as possible with the aid of data mining and data warehousing technologies. |
| **Research questions** | 1. How to make Data Analysis and Reporting process Automated instead of manual? 2. What is Data Reporting & Analysis are meant for? 3. Why Data Mining & Reporting are keep gaining popularity in today’s day-to-day activities? 4. What are the needs of data mart or data warehouse for Business Intelligence? |
| **Data Collection Technique** | Qualitative – Interview  Quantitative – Survey |
| **Analysis & Projection** | Basically, data mining is of two types supervised and unsupervised learning techniques. But this research confine itself in using supervised learning technique, wherein the algorithm is trained by feeding it with data for making prediction.  But however in the near future it could be improve and updated for the implication of unsupervised learning technique wherein the algorithm will lean on it own. |

* 1. **Justification of Methods**

Well, in this research project I used both qualitative and quantitative research to conduct my data gathering from correspondent.

* 1. **Tools and Techniques**

Technology For development or coding, there are were two difference programming languages used:

Plain/SQL – Structured Query Language, was used as native language for designing, querying and managing database.

Java programming language – Java swing was used for designing user interface to interact with the database for generating report.

Development tools:

There are two development tools that were used:

NetBeans 8.2 Integrated Development Environment- this was used for the development portal wherein Java Swing was used event handling of the user interface.

Oracle Relational Database Management System – Was used as database management system for designing, querying and managing database.

Services and Interfaces:

Java swing is used as interface

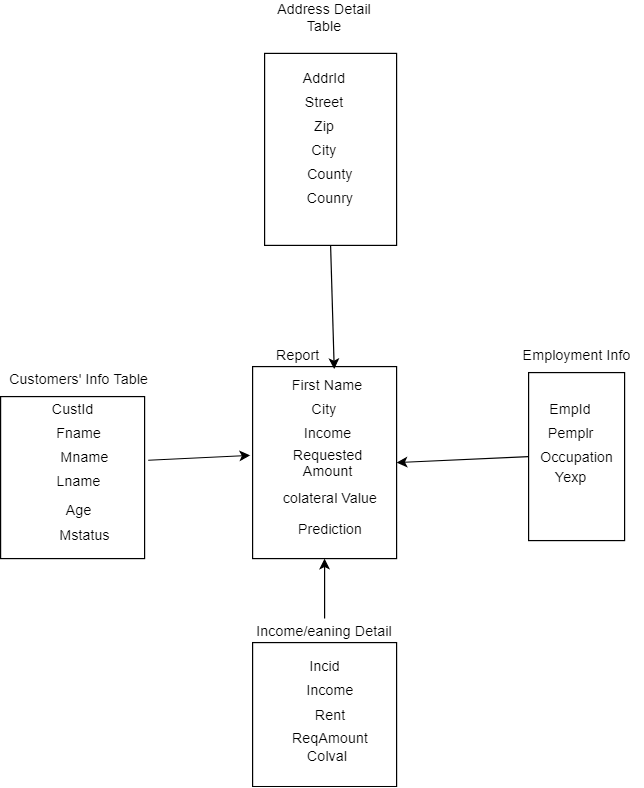
1. **Data Analysis & System Design**

Indeed, with the interview conducted with the loan officers at International Bank Liberia Limited. I learnt that the analysis of loan applicant is been done manually in reviewing their records as well as their assets.

So base on that, this system will take in consideration of customer’s basic information, customer’s income detail, customer address detail, employment information. And with these information there will be a template for generating a report.

**Design:**

The diagram below depict the design of Decision Support Portal System.



**Explanation of the Data Dictionary of the design:**

1. Address Detail table – This is the table where information about loan applicant can be found. It stores address information of a bank’s customers.

Basically it has six entities, namely:

Addrid – this is the unique id assign to a customers.

Street – this stores customer’s street or residential address name.

Zip – this attributes stores zip code of the customer’s residential area.

City – save the name of the city.

County – this attribute store the name of the county for city in which the customer live.

Country – store the name of the country in which the customer live.

1. Customer’s Information Table – this table is one of the importance table, it store information about a given customer. Basically, their particular.

It has six column, namely:

CustId –this column store the auto-generated customer’s id.

Fname – this column store the first name of the customers.

Mname –this column save the middle name of the customers if available.

Lname –this column save the last name of the customers.

Age –column stores the calculated age of the customers.

Mstatus –store the customer’s marital status.

1. Earning/Income detail –this is a table where information about customer’s income are entered in the operational database, oracle.

.This table has five column:

Incid – this is the auto-generated id assign to each customers.

Income –this column store the income level information of the customers.

Rent – this column save the liability of a customers.

ReqAmt –this is the requested amount that a customers or clients a might apply for as a loan.

Colval –this column store the collateral value that a loan applicant might present to the officer.

1. Employment information – provides information about customer employment information. It has four columns:

EmpId –this column save the customer employment IDs

Pemplr – this column save the present employer of the customers

Occupation –this column save the occupation of the loan applicant if unknown.

Yexp –year of experience, this column store the year of experience of the applicant or the customer.

1. Report table –this is the virtual table that fetch data from these four tables and generate it as report. It has six columns that will be used in the report:

First Name –this column fetch the first name of the customer from the operational database.

City – this column display the name of the city of the applicant.

Income –generate the income of the customers

Requested Amount – this column generate the requested amount from operational data source

Colval –collateral value, this column fetch collateral value from the system.

Prediction –this column give the prediction on customers payment behavior as “Safe” or “Risky” base on specify condition.

* 1. **Feasibility Study**

This analysis takes into consideration as weather it is technically, economically and operational feasible to undertake this project assuming if it was meant for production environment. At the end of this study, I will get to determine the possibility of the development of the system.

* + 1. **Technical Feasibility Study** –this aspect of the study specify hardware equipment and software that will successfully support the task.

Considered hardware equipment include the following:

Client/Server environment fully setup

Server Machine for hosting the application

Available client machine for accessing the application

Considered compactible software include the following:

On Windows host, Windows server should be installed and running

On Linux host, Ubuntu server should be installed

On Windows or Linux host, a compactible java run-time environment should be installed as well as NetBeans Integrated Development Environment.

Again on the host server, Oracle Database should be installed, version should be less than 11g.

On client-side, a copy of the client application should be installed along with Java run-time environment.

* + 1. **Economic Feasibility Study**

**This will take in consideration of the famous cost benefit analysis.**

|  |  |
| --- | --- |
| **Cost** | **Benefit** |
| Necessary available hardware and software products | Economically feasible |
| High service charge from java developers | Platform independent |

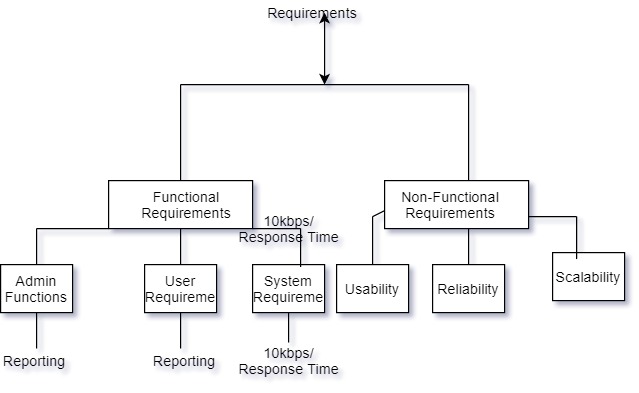
* + 1. **Operational Feasibility**

This aspect of feasibility study focuses on people, organization and economy.

The diagram above depict the study:

|  |  |
| --- | --- |
| Technical Performance | System Acceptance |
| Well, with the aid of embedded algorithm, the system will be able to provide the necessary report to users without any hindrance. | In term of user interaction with the system in generating, there resistance will be less. |

1. **Requirement Planning he diagram shown the necessary requirements:**

****

* 1. **Functional Requirement**

The functional requirement of the system takes into consideration of the following functions:

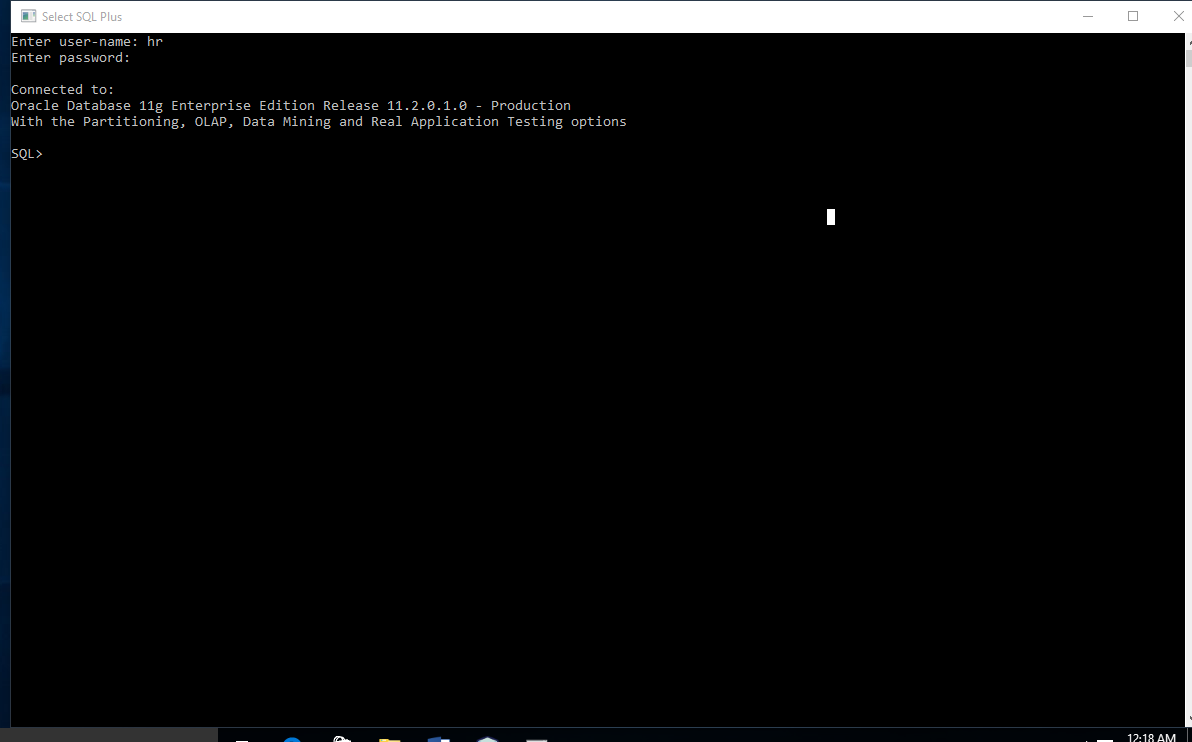
1. Administrative functions: these are the routine things the system will do, such as user management and data reporting.
2. User requirements –these are what the user of the system can do, such as generating report.
3. System requirements –these are things like software and hardware specifications, system responses, system actions.
   1. **Non – Functional Requirement**

The functional requirement focuses on the following:

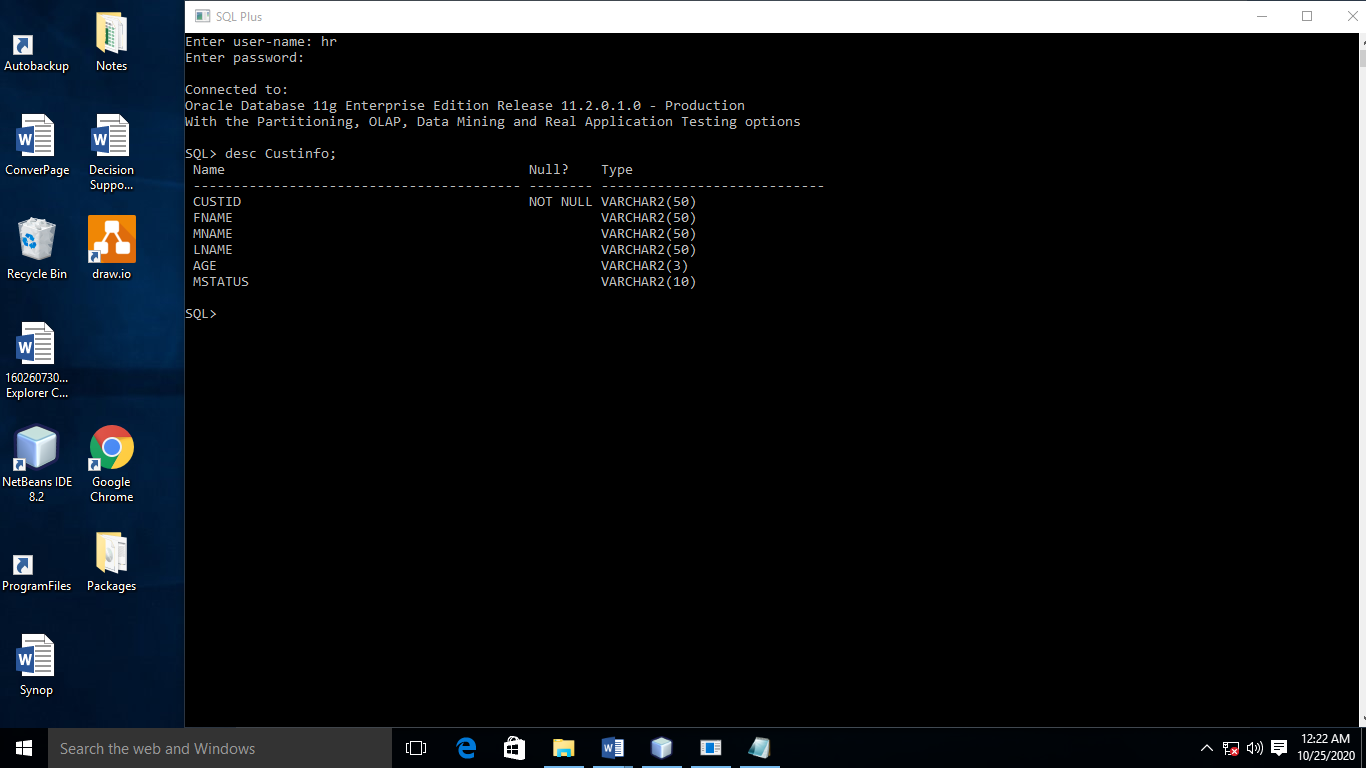
1. Usability –this focuses on the appearance of user interface and how people interact with it.
2. Reliability –this non-functional requirement focus on the uptime of the system. Normally, on that it depends on the type of the server hosting the application
3. Scalability –this non-functional requirements focuses on the proper operation of the system in case of future change.
   1. **System Implementation (Screenshot)**

For the implementation of this system, first, network computer that will be allow to access the application. And this is because the system on shared network (extranet). The make sure java run time environments are installed on user computer. When that is done, a copy of the application can be installed on individual computer.

The screenshot below depict plain sql/oracle login interface in order to access the Operational Data Source for further data entry.

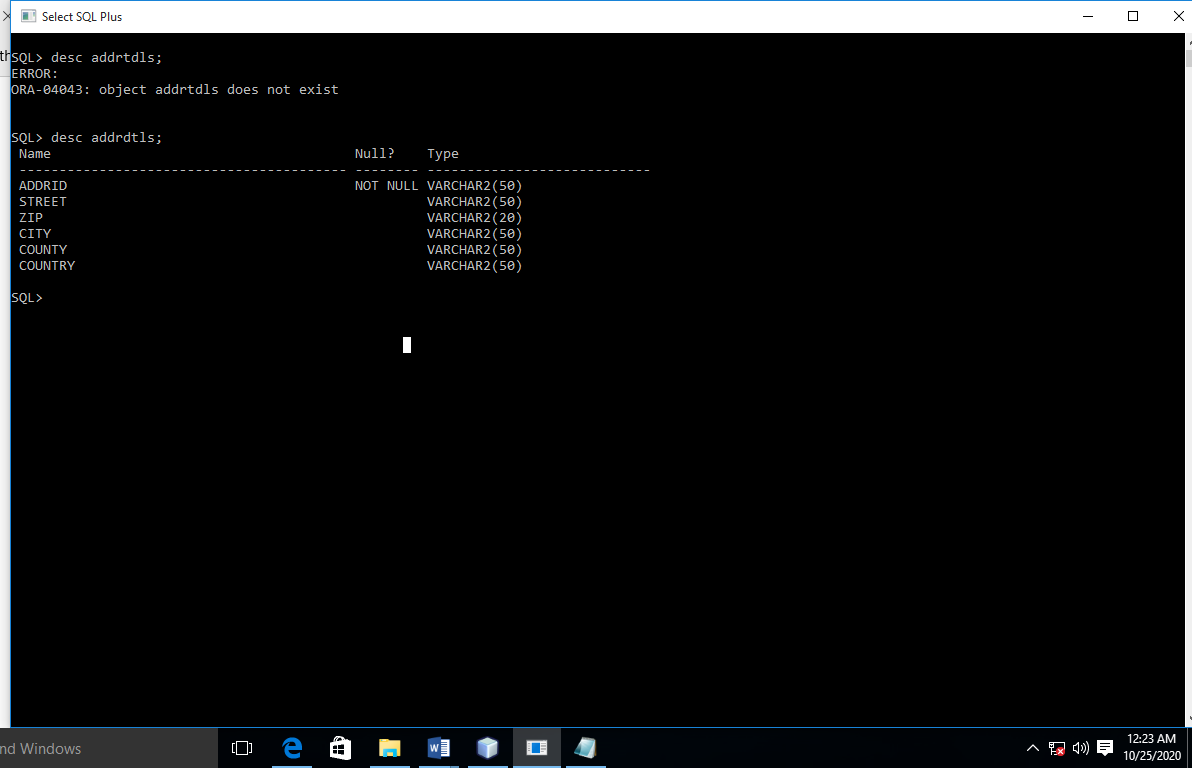


The screenshot below depict the table description for Customer information at the operational data source.

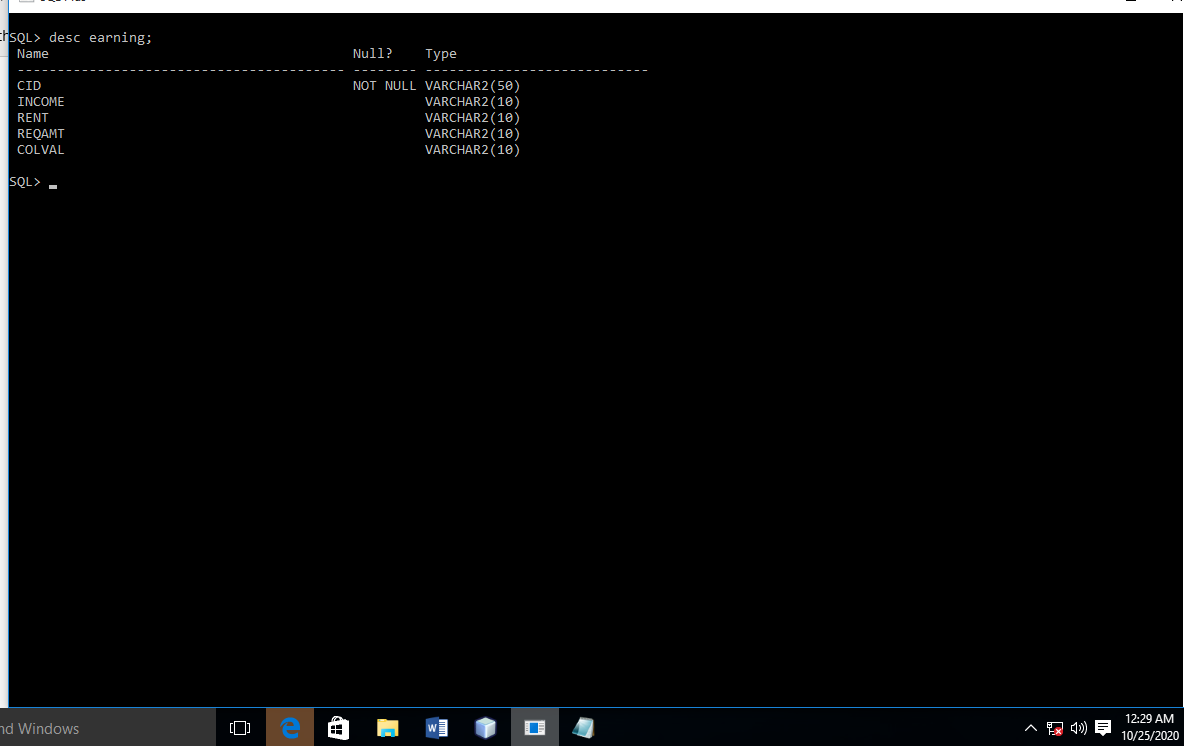


The screenshot below depict the table description for customer address detail at the

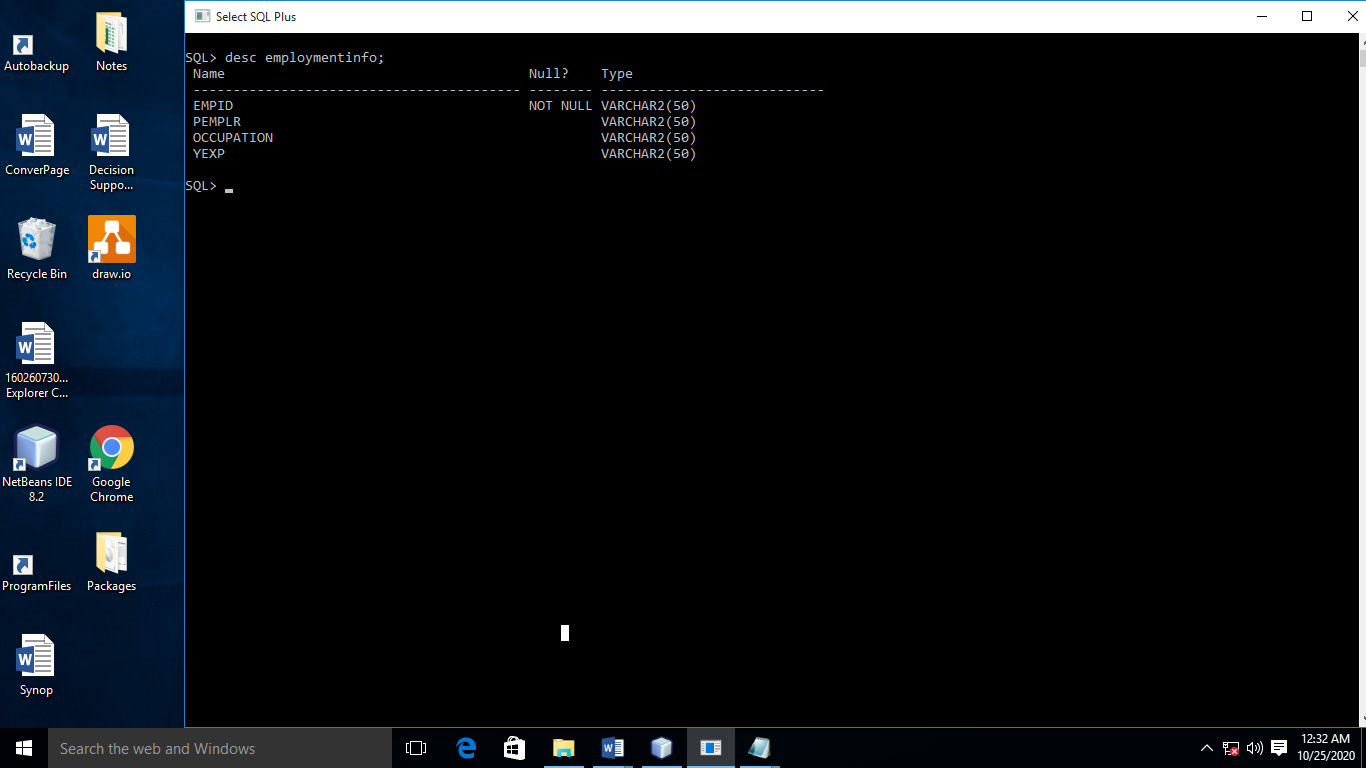
Operational data source.



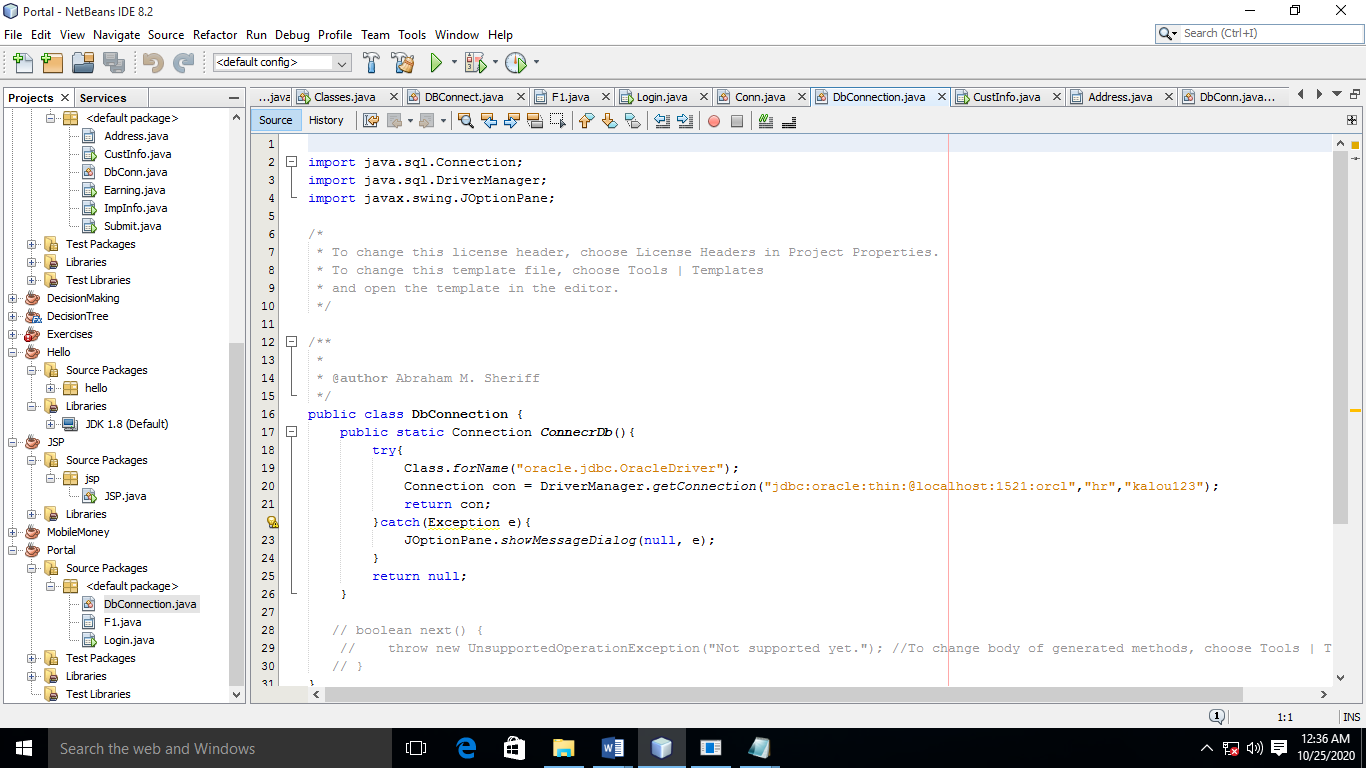
The screenshot below depict table description for customer’s earning or income detail at the operational data source.



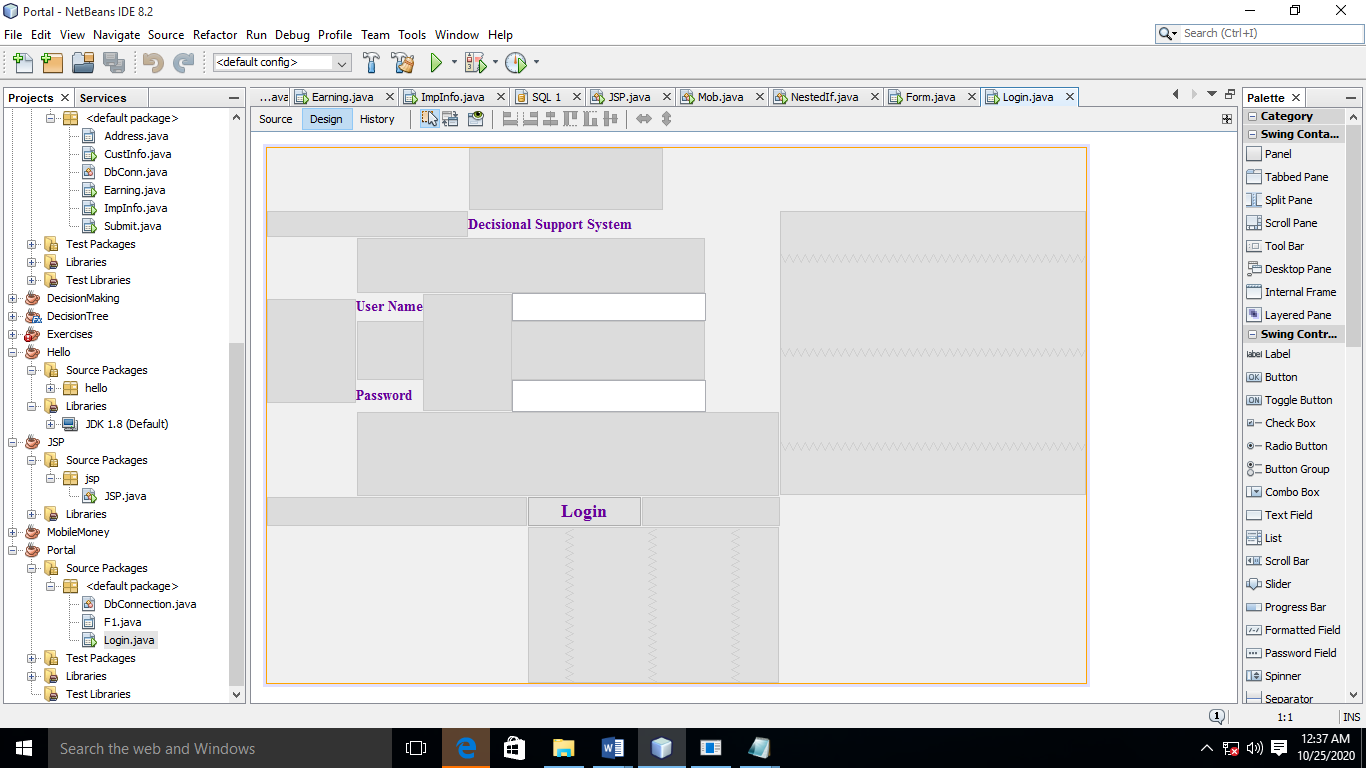
This table depict the operational data source operational data source for customer or client employment info.



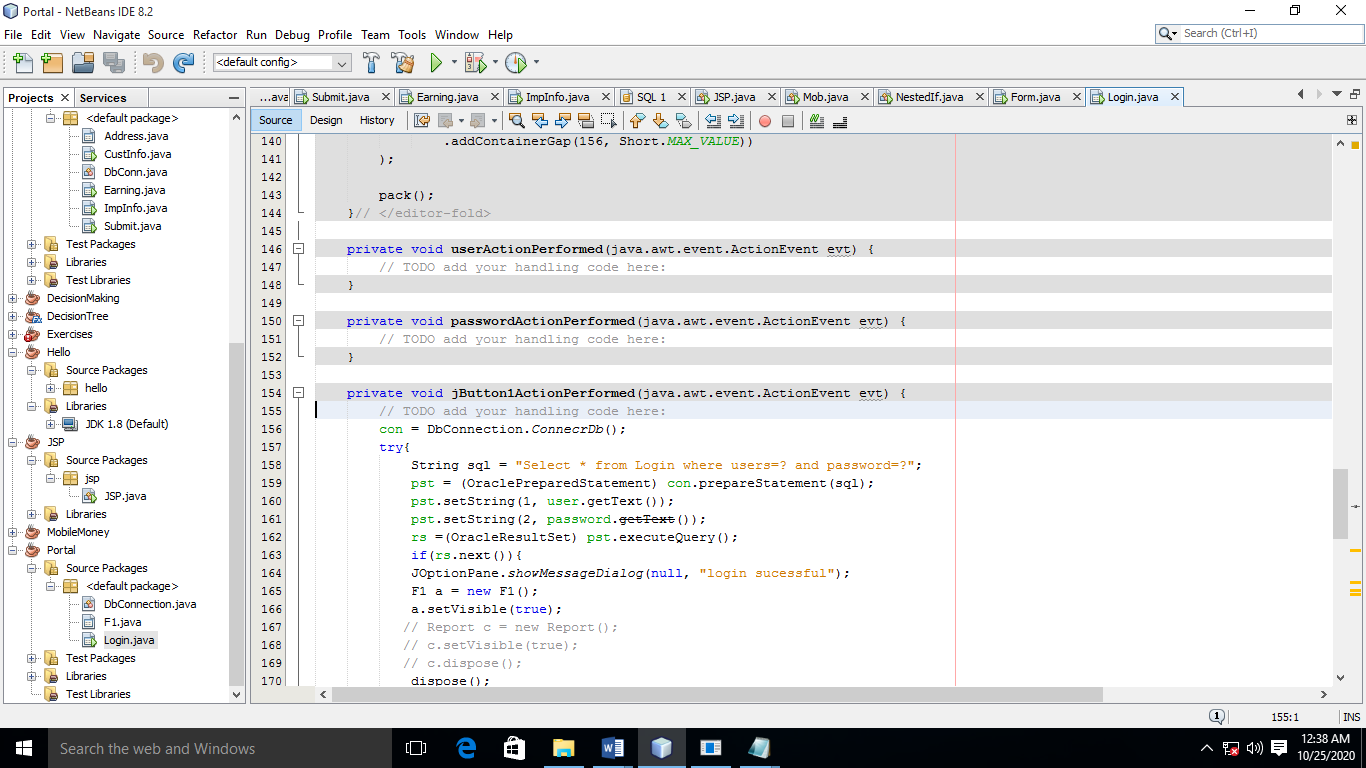
DB connection class in java to operational data source (Oracle). For the successful connection to oracle, I imported ojdbc.jar library to the application packages.



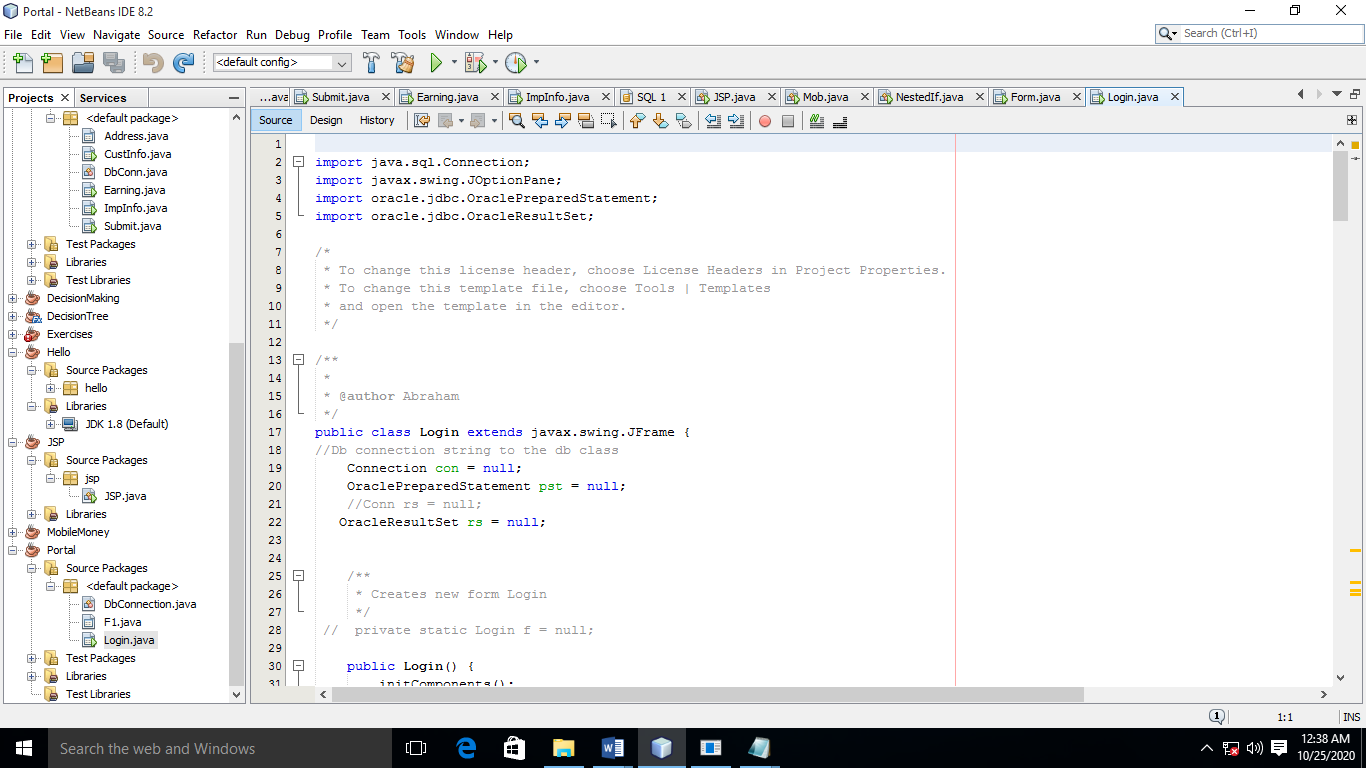
Login interface form for end-users. This form is auto-generated when you create a window within the prefer package of the project.



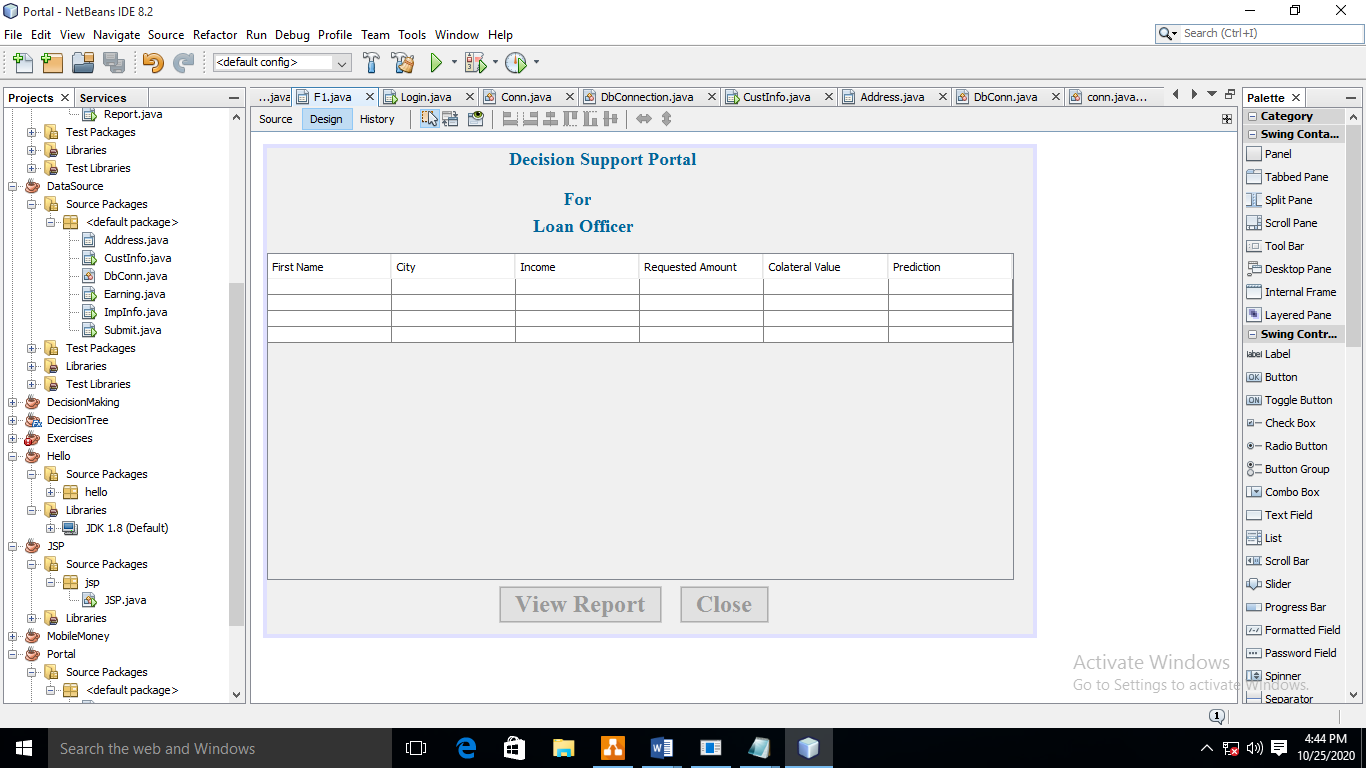
Linking the login and report form to the database class.



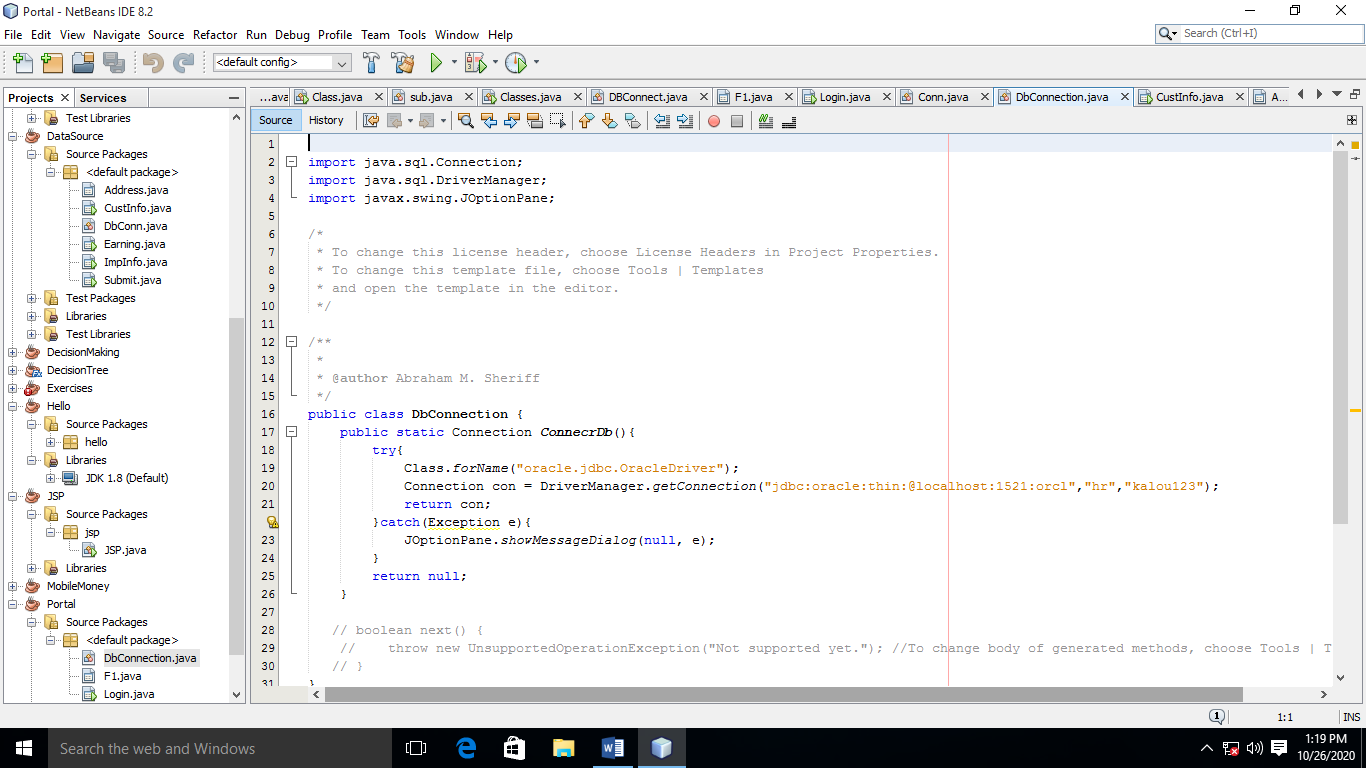
Import driver for calling libraries for DB Connection.



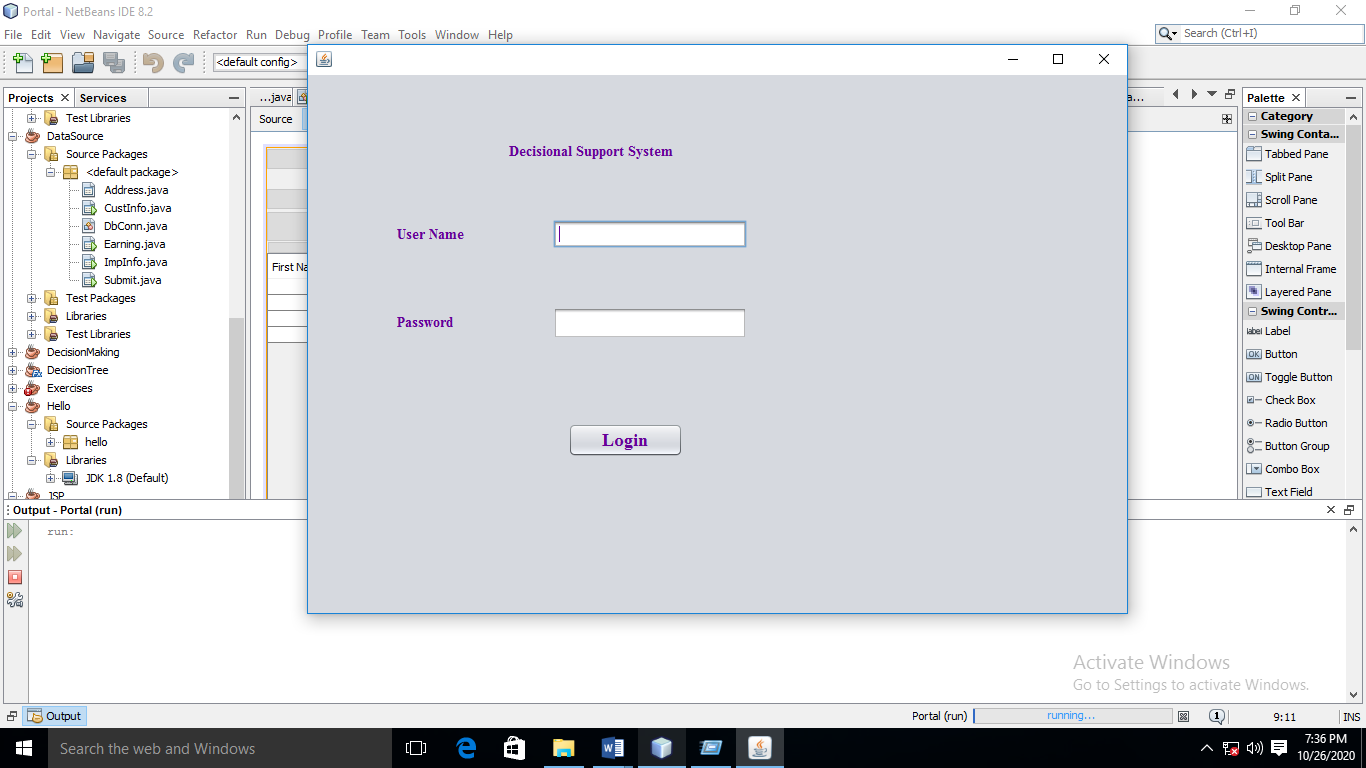
Report template interface form, is where report will be generated for intelligence decision making.



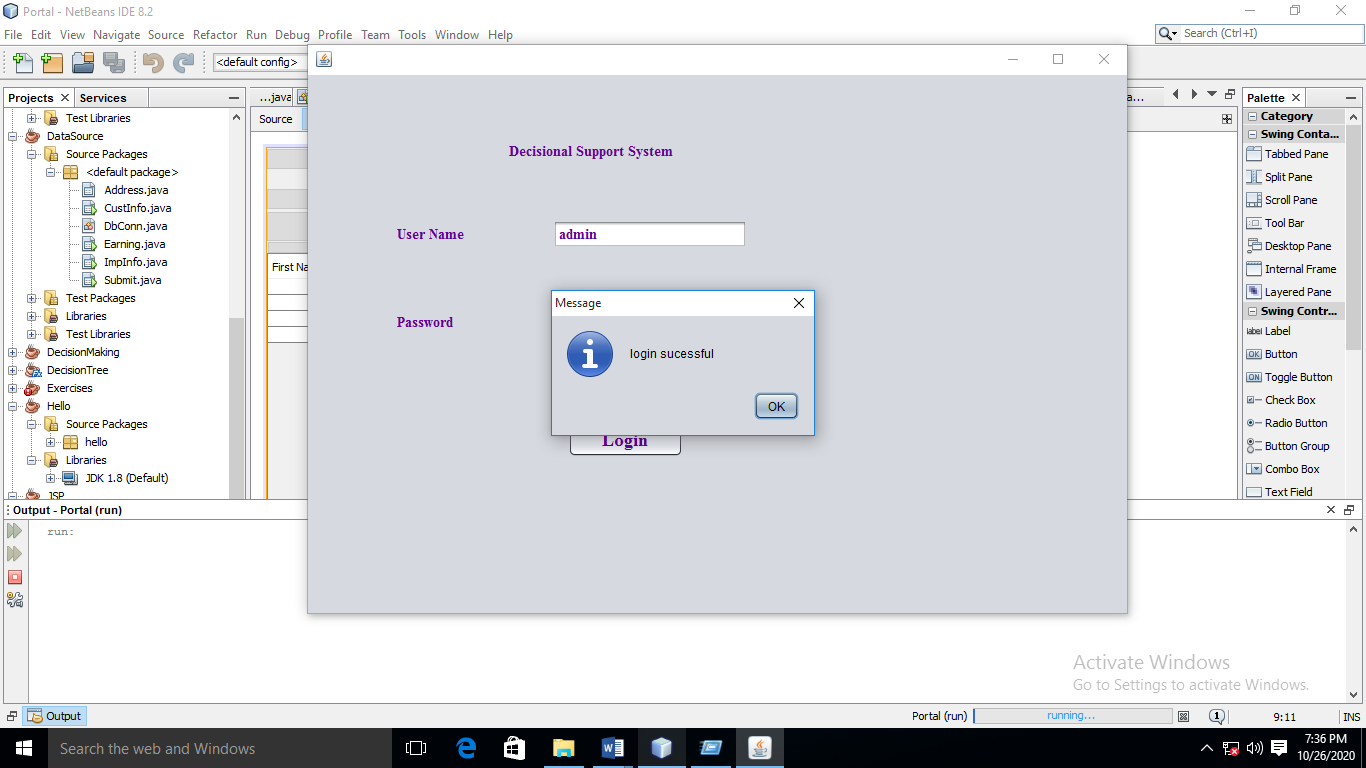
The screenshot below depict the connection to oracle database using jdb oracle driver with thin API



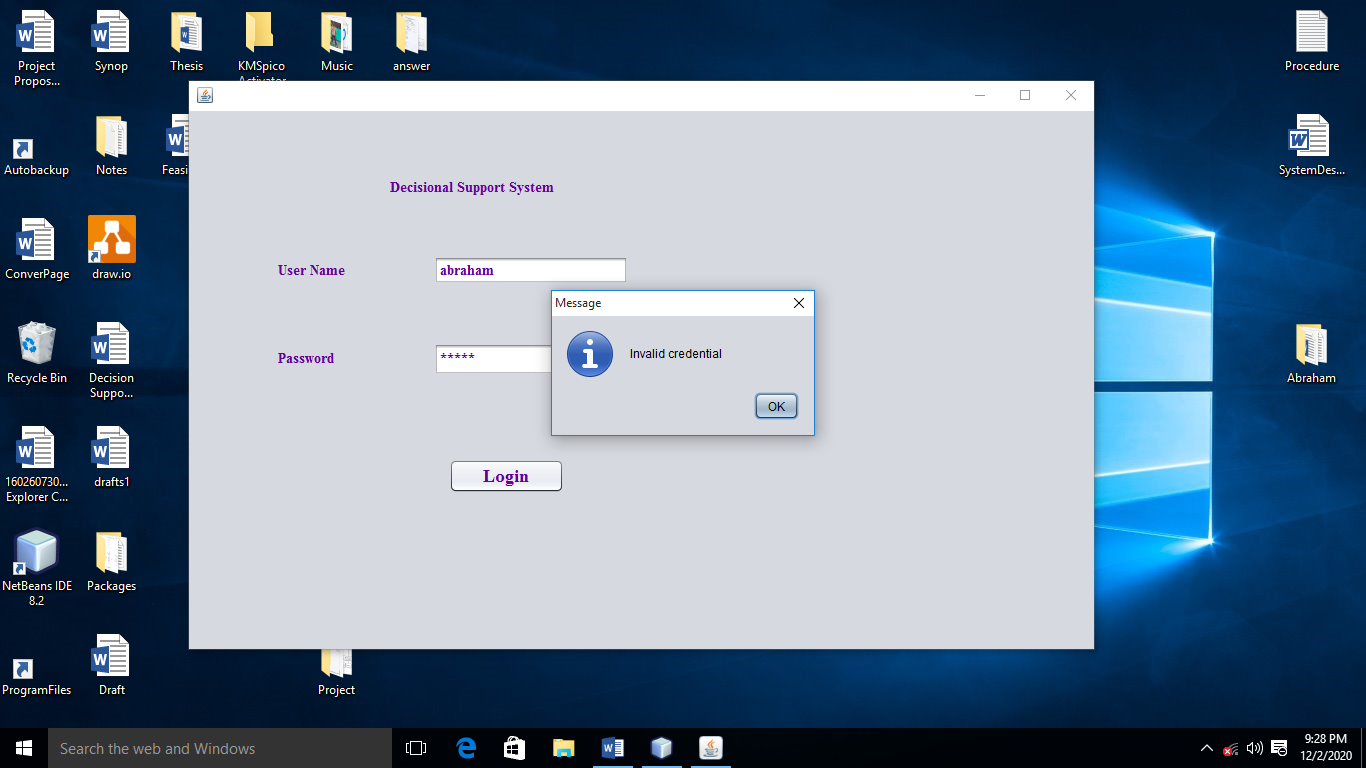
The screenshot below indicate login interface of the system. If the username and password is correct, the attempted user will be able to access the portal or else access will be denied with warning and exit the App.

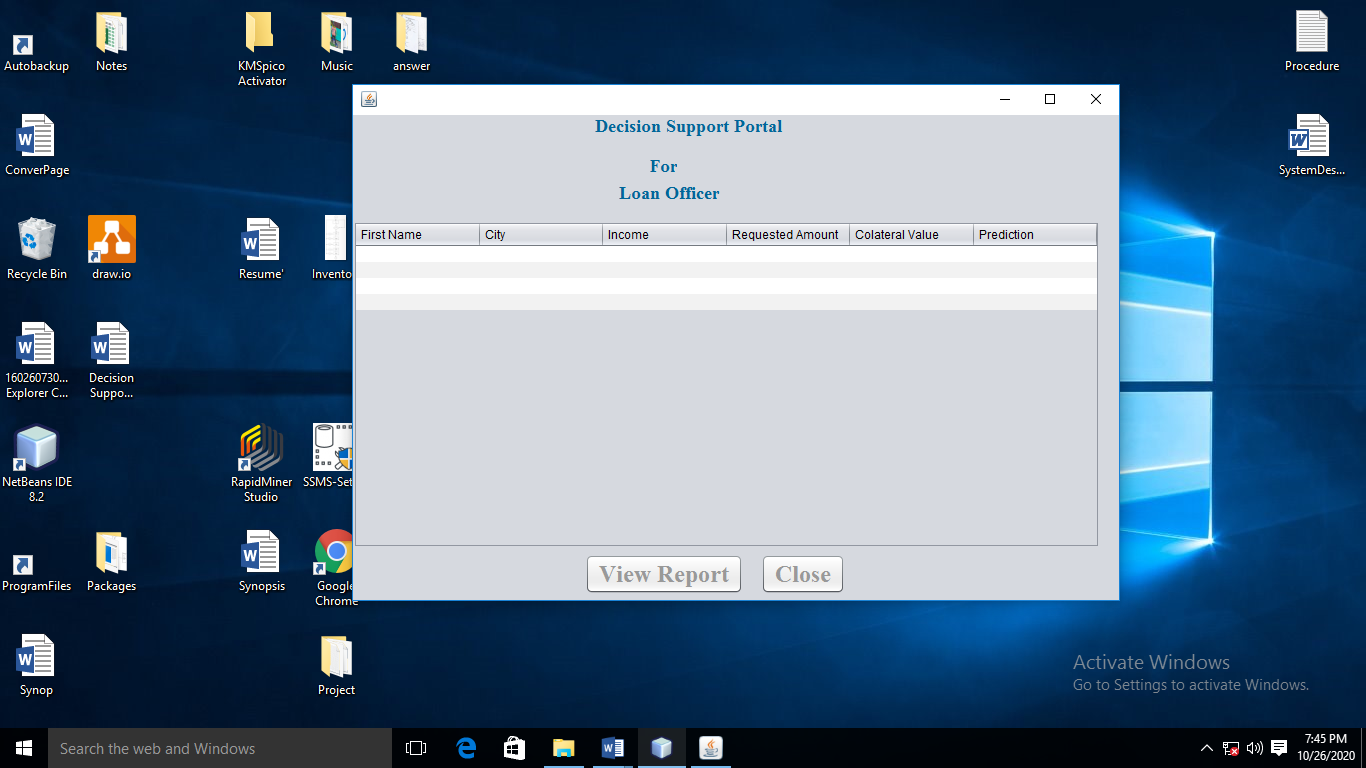


The screenshot below depict a successful logon to the portal after which a user has entered his or her login credential.

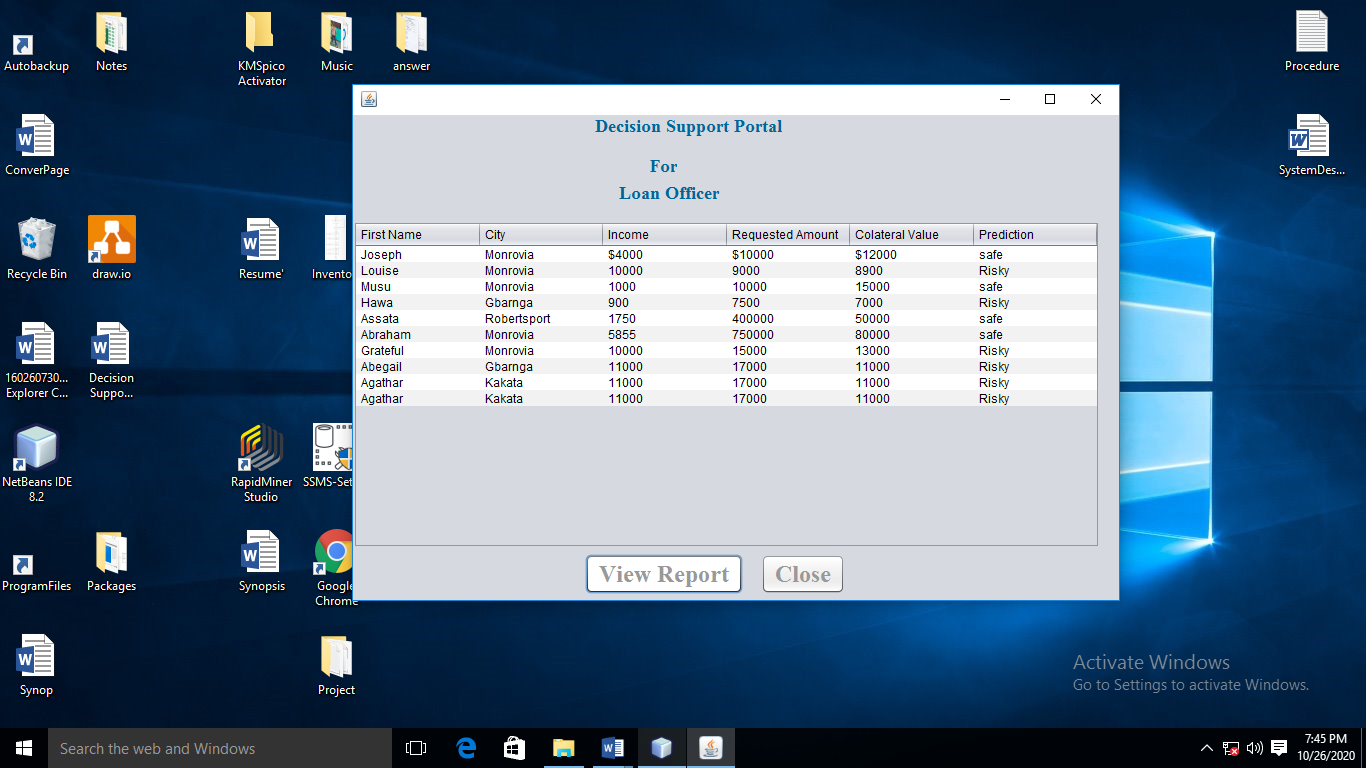


The screenshot below indicate unsuccessful login to the portal.



The below sreenshot show report template for the system and is from there report are generated through View Report button 

The diagram below represents a partial report from the system



* 1. **System Testing**

The chart below depicts the test cases for Decision Support Portal. Cases take in consideration of successful logon, unsuccessful logon, report generation and closure of the system normally.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Test Case Scenario | Test Steps | Test Case Description | Test Data | Expected Result | Actual Result | Pass/Fail |
| Check users successful logon | Locate the application your system and double click on it. | To check whether a valid user name and password was entered. | User Name = admin  Password = admin@123 | Login should be successful | Login was successful | Pass |
| Check users unsuccessful | Locate the application your system and double click on it. | To check whether an a unauthorized login fail | User name = Abraham  Password=Pass@123 | Login Should be unsuccessful | Logon was unsuccessful | Fail |
| Ensure successful Report generation | Double click on “View Report “ button upon displaying the applicant | Check whether a “View Report” button generate report | Click on “View Report” should display report. | Should display report | Should not display report | Pass |

* 1. Conclusion & Recommendation

Primarily, decision support portal promises to generate pattern from collected data that will be gathered from operational data sources and predict the behavior of loan applicants in bank as either “Risky” or “Safe” base on the person income, age, collateral value and requested amount.

Furthermore, in order to fully accomplish this task, Knowledge Discovery in Databases (KDD) process have to be performed. Wherein data will firstly be collected, from there carry on data integration and cleaning. Secondly there will be data selection and transformation. Thirdly, do the mining on the data. Follow by pattern evaluation and knowledge presentation will be carried on.

And again, after KDD process, the Extract, Transform and Load (ETL) process will be proceeded for loading data from operational data sources to data warehouse, where complicated SQL queries will be written for generating data or report.

With the advancement in technology today, I envisage tremendous change in the following technology.

**Multimedia Data Mining**

This is one of the latest methods which is catching up because of the growing ability to capture useful data accurately. It involves the extraction of data from different kinds of multimedia sources such as audio, text, hypertext, video, images, etc. and the data is converted into a numerical representation in different formats. This method of data mining can be used in clustering and classifications, performing similarity checks, and also to identify associations.

**Ubiquitous Data Mining**

Ubiquitous Data Mining is the process of extracting hidden classifier, cluster, frequent item-set, and association rules from distributed data among a number of mobile and stationary data sources.

This method involves the mining of data from mobile devices to get information about individuals. In spite of having several challenges in this type such as complexity, privacy, cost, etc. this method has a lot of opportunities to be enormous in various industries especially in studying human-computer interactions.

**Distributed Data Mining**

This is another future scenario in technologies, as the world continues to undergo a rapid change in technologies particularly in the area of Artificial intelligence, Machine Learning and Internet of Things.

This type of data mining is gaining popularity as it involves the mining of huge amount of information stored in different company locations or at different organizations. Highly sophisticated algorithms are used to extract data from different locations and provide proper insights and reports based upon them.

**Spatial and Geographic Data Mining**

This is another new trending type of data mining which includes extracting information from environmental, astronomical, and geographical data which also includes images taken from outer space. This type of data mining can reveal various aspects such as distance, is 0omainly used in geographic information systems and other navigation applications.

**Time Series and Sequence Data Mining**

The primary application of this type of data mining is study of cyclical and seasonal trends. This practice is also helpful in analyzing even random events which occur outside the normal series of events. This method is mainly being use by retail companies to access customer's buying patterns and their behaviors.

In addition to that, in order for data to really be valuable to an organization, we need to be able to discover patterns and relationships within that data. And that is what data mining does. Those connections and insights can enable better business decisions. Data mining can also reduce risk, helping us to detect fraud, errors, and inconsistencies that can lead to profit loss and reputation damage. Different industries use data mining in different contexts, but the goal is the same: to better understand customers and the business. Nowadays, data mining is widely use in area like E-ecommerce, education, Telecommunication, banks, healthcare for generating report and intelligence decision-making. The first example of Data Mining and Business Intelligence comes from service providers in the mobile phone and utilities industries. They organize billing information, customer services interactions, website visits and other metrics to give each customer a probability score, then target offers and incentives to customers whom they perceive to be at a higher risk of churning.

**Retail:**

Another example of Data Mining and Business Intelligence comes from the retail sector. Retailers segment customers into ‘Recency, Frequency, Monetary’ (RFM) groups and target marketing and promotions to those different groups. A customer who spends little but often and last did so recently will be handled differently to a customer who spent big but only once, and also some time ago. The former may receive a loyalty, upsell and cross-sell offers, whereas the latter may be offered a win-back deal, for instance.

**E-commerce:**

Perhaps some of the most well-known examples of Data Mining and Analytics come from E-commerce sites. Many E-commerce companies use Data Mining and Business Intelligence to offer cross-sells and up-sells through their websites. One of the most famous of these is, of course, Amazon, who use sophisticated mining techniques to drive their, people who viewed their product, also liked this functionality.

* 1. Key Contribution

No tasks is single man’s effort. It is done through cooperation and coordination of other people. Based on that, I would like to acknowledge the contribution of the following personalities for their tireless effort in guiding me for carry on this project successfully.

Mr. Shyam Nair –head of Department for Information Technology (IT), BlueCrest University College Liberia

DR. Kennedy – Adjunct Lecturer, BlueCrest University College Liberia

* 1. **Limitation & Future Work**

With regard to the diversity of Data Science, the study will not be generalizable to all area of data mining and data warehousing. Also the finding could be subject to a lot of interpretation.

Data mining, knowledge discovery, or predictive analysis – all of these terms mean one and the same. Broken down into simpler words, these terms refer to a set of techniques for discovering patterns in a large dataset. These patterns help in creating a predictive model to stay on top of the future behaviors.  
Today, most of the organizations – irrespective of their domain – are looking to capitalize on their Big Data and are hence using sophisticated analytical methods. As the consumption of Big Data grew, so did the need for data mining. Today, we can see examples of data mining everywhere around us.

The use of Data Mining and Analytics is not just restricted to corporate applications or education and technology, and the last example on this list goes to prove the same. Beyond corporate organizations, crime prevention agencies also use data analytics to spot trends across myriads of data. This data includes information including details of all the major criminal activities that have happened.  
Mining this data and thoroughly studying and understanding patterns and trends allows these crime prevention agencies to predict the future events with much better accuracy. With the help of Data Mining and analytics, these agencies can find out everything from where to deploy maximum police manpower (where is the next crime most likely to happen and when?), who to search at a border crossing (based on type or age of the vehicle, number or age of occupants, or border crossing history), to even which intelligence to take seriously in counter-terrorism activities

Today, data mining plays a pivotal role in implementing data reporting and analysis for intelligence decision making process. And it is mainly applicable with the following services or products.

**Service Providers**

Service providers like Telecommunication, insurance and bank have been using Data Mining to retain customers for a very long time now. Using the techniques of Business Intelligence and Data Mining allows these service providers to predict their customers’ behavior

Today, every service provider has terabytes of data on their customers. These data includes things like your billing information, customer services interactions, website visits, and so on. Using mining and analysis of these data, the service providers assign a probability score to each customer. This probability score is a reflection of how likely you are of switching the vendors. Then, these companies target the people at a higher risk by providing incentives and personalized attention, to retain the customers.

**Supermarkets and Retail Stores**

Also, Data mining allows the supermarket owners to know your choices and preferences even better than yourself. If you don’t believe it, you’ll be amazed by what target did a few years back.   
Following the purchase history and behaviors of one of their female customers, Target correctly concluded that she is pregnant. Oh, and let’s tell you – this was even before the woman herself knew. Such is the power of data, patterns, analysis and association rule mining.   
In general, these retail stores divide the customers into what they call “recency, frequency, monetary” (RFM) groups and specific groups with different campaigns and strategies. So, a customer who spends a lot but infrequently will be dealt differently than a customer who spends little but often. The latter kind may receive loyalty, upsell, or cross-sell offers, whereas the former might be offered a win-back deal, just for instance. This occur most often with the implication of Association Rule Mining.

* 1. References

Ms. Kushboo, Dr. Akash Saxena, Mr. Sandeep Sexana – Data Mining and Data Warehousing

Sumit Prakash Tayal – Software Project Management

Ivan Ben Ross – Relational Database Design

Shosbana and Maheshwari and Savithri - “Study on Big Data with Data Mining” April 2015,

Sauravkaushik,- “An Introduction to Clustering and Different Methods of Clustering” November 2016

Dr. Madhulika - Structuring System Analysis and Design

Dr. K. Chandrasheka - Software Engineering

Shashi Sign Management Information System

* 1. Appendixes

Author’s Name: Ivan Bayross

Title: Relational Database Design

Book Chapter: 5

Topic: Query Optimization and Performance

Footnote: Oracle’s approaches to optimization – rule-based and cost-based approach

Author: Sumit Prakash Tayal

Title: Software Project Management

Book Chapter: 3

Topic: Software Project Planning

Sub-topic: Constructive Cost Model (COCOMO)

Footnote: Using Constructive Cost Model for estimating the cost of a project.

Author: Ms. Khushboo Saxena

Title: Clustering Technology

Chapter: Cluster Analysis

Topic: Neural Network Approach in Clustering Technology

Footnote: