UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

Application of Data Mining in Pharmaceutical Imports in Kenya

BY

Victor Wahinya Kahonge
P56/72700/2012

Supervisor
Mr. Christopher Moturi

May 2014

Submitted in Partial Fulfilment for the Award of the Degree of Master of Science in Information Systems of the University of Nairobi

DECLARATION

STUDENT

I, the undersigned, declare that this project is my original work and that it has not been presented in any other University or Institution for academic credit.

NAME: VICTOR WAHINYA KAHONGE REG. NUMBER: P56/72700/2012

Signature: __________________________

SUPERVISOR

This research project has been submitted for examination with my approval as University Supervisor.

Signature: __________________________ Date: _________________________

MR CHRISTOPHER MOTURI

SCHOOL OF COMPUTING AND INFORMATICS

UNIVERSITY OF NAIROBI

ABSTRACT

Data mining which is often referred to as Knowledge Discovery in Databases (KDD) is a sub-discipline of computer science aiming at the automatic interpretation of large datasets. Its end result is the extraction of meaningful patterns from data to aid in the process of decision making. Data mining is a discipline that is applicable to all subject areas from aviation to zoology as most have transaction data.

Historical data from permits prior to the proposed system is of high value regardless and is of great interest to the study as it proposed to enhance usability of such data for decision purposes. All data on imports and exports available electronically was captured through the data capture system. Currently, PPB uses regular queries in its databases to generate reports. These reports only present one facet of data without presenting the true picture to the concerned decision makers leading to inaccurate decisions.

Objectives of the study were to analyse data on imports and exports of pharmaceutical products in Kenya and discover patterns of association and correlation between the various pharmaceutical product groups.

The study adopted the CRISP-DM framework for data mining and utilised RapidMiner as a tool for data analysis and mining. CRISP-DM is a generic framework that is applicable in most subject areas and is quite tested. RapidMiner is an open source data mining tool that is quite popular world-wide due to its capabilities, ease of use and availability of online help. This study applied data mining to the field of pharmacy regulation. The study analysed data on imports of pharmaceutical products for interesting patterns. The study performed correlation and association analysis on the import data.

The study proved that there exist patterns in pharmaceutical import data. The patterns are also similar to prescription patterns from studies in Ethiopia, Nigeria and India especially on association of several pharmaceutical product groups.

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.0. Background

A medicine is any drug or substance used to treat and or prevent disease. According to the Pharmacy and Poisons Act, a drug or therapeutic substance is anything that may cause effects including but not limited to medicinal, intoxication and performance enhancement when ingested by a human or an animal and is not a food. The Pharmacy and Poisons Board (PPB) is mandated by the Kenya Government to regulate the import and export of all medicines and their raw materials which are at times referred to as Active Pharmaceutical Ingredients (API).

Most organisations have implemented or are in the process of implementing information systems due to the need for automation. This is mainly due to the ease of updating records into repositories or databases and the high speed of retrieval when historical records are required. Therefore information systems are valuable tools for efficient transaction processing and data capture during their operations. Over time, the data accumulated by organisations in their databases becomes vast and can be used for gain if the right people have the correct tools to properly analyse it and give intelligent conclusions from it. In this day and age, the accumulated data reserves created by organisations are very valuable and could be used to predict future trends, patterns or possible pitfalls to an organisation. Such vast data reserves can be termed as “Hidden Treasure Chests” which are at times impossible to utilise (Kumar, Sehgal, Sehgal, & Chauhan, 2012).

Without advanced methods of analysis and examination, these vast amounts of data may be rendered useless. “Data mining” which is often referred to as “Knowledge Discovery in Databases” (KDD) is a young sub-discipline of computer science aiming at the automatic interpretation of large datasets (Kriegel, Borgwardt, Kröger, Pryakhin, Matthias, & Zimek, 2007). Data Mining, a relatively new field of analysis, is defined as the process of discovering several models, summaries and derived values from a given collection of data (Kantardzic, 2011). It incorporates the application of Statistical and machine learning methods to obtain solutions to current problems. Previously, Statistical methods have been in use to analyse vast amounts of data although currently, Relational Database Management Systems (RDBMS) come equipped with modules that perform Data mining that is more advanced. Such RDBMS include Oracle and

Microsoft SQL Server. Vast amounts of data present from various sections in an organisation are housed in a Data Warehouse. The Pharmacy and Poisons Board (PPB) as a regulatory authority in Kenya has put in place systems which generate vast amounts of data that is quite valuable if exploited with the correct tools so as to enable management unlock the relevant knowledge lying hidden in its databases and files. Currently, the Board does not employ any tools relating to artificial intelligence in the extraction of knowledge from its data. However, with the current technological advances and the need to adapt to current trends will move it to utilising the data they hold for more than what they are using it for.

1.1. Problem Statement

The importance of this research is indicated by the fact that PPB, like any other organisation in the current digital age, is in the process of digitising its historical data on imports and exports of pharmaceutical products. Such data is quite valuable to management and decision makers in that it holds valuable information that could influence policy that would in turn benefit the Kenyan citizens in the long run. PPB is in a transition period whereby a new transaction processing system is in the process of implementation which will be capable of collection of data from permits on imports and exports and consequently, they will have considerably clean data for analysis.
Historical data from permits prior to the proposed system is of high value regardless and is of great interest to the study as it proposed to enhance usability of such data for decision purposes.

All data on imports and exports available electronically was captured through the data capture system. Manual data entry was done for all data on imports and exports. Manual entry data is prone to many errors (Rahm & Do, 2000). For an import or export to take place, an agent needs to apply for a permit from PPB. The permit may contain one or many products depending on the need of the particular agent. The current technological advancements especially in Artificial Intelligence and in particular in the realm of Knowledge Discovery can be put into great use in such an organisation as PPB. We believe if incorporated, PPB might unlock valuable information from their databases that may enable make informed decisions in terms of policy, economy and public interest. Currently, PPB uses regular queries in its databases to generate reports. These reports only present one facet of data without presenting the true picture to the concerned decision makers leading to inaccurate decisions.

This study will benefit PPB and its stakeholders whose mandate is public service and regulation. The analysis of pharmaceutical products import and export in Kenya is a very sensitive issue and if their regulation is done effectively and accurately, then there will be smooth operations. PPB may also acquire useful information to use in its regulatory action.

1.2. Research Objectives

The following are the specific objectives of this study

1. To analyse data on imports and exports of pharmaceutical products in Kenya
2. To discover patterns of association and correlation between the various pharmaceutical product groups

1.3. Justification

Although there have been technological advancements in Artificial intelligence that have radically surpassed the capacity and performance of statistical methods in reporting and extraction of knowledge from databases, the level of adoption of knowledge discovery methods in Government institutions (including PPB) is wanting. PPB is a very critical Government institution charged with the monitoring of all pharmaceutical Products in the country. Data mining adds value to a Data Warehouse. Any tool or model that will be of use in the efficient extraction of vital information on the trend and movement of such products should be an asset to it.

Pharmaceutical products are highly sensitive in terms of monetary value and social implications as some of them are utilised in the black market to produce drugs which can be abused by the general public. Other pharmaceutical products such as antibiotics are prone to misuse leading to drug resistance (Afsan, Haque, Alam, & Noor, 2012). Such drug abuse is much prevalent in our youthful generation. It is a mandate of the Government to keep its citizens safe through the control of such products and this mandate is carried out by the Board. Statistical techniques alone cannot analyse such data for trends, patterns and reports. This study proposed to provide a solution that will enable the utilisation of import and export data held by PPB.

Based on the proposed output of the project, the benefits of obtaining patterns of the various business entities as either importers or exporters will be very valuable to the Board as it will better equip them in the process of making decisions and forecasting future needs.

1.4. Scope

The project entailed the entire process of knowledge discovery within the databases housed in PPB. The project was carried out in the PPB premises. The study utilised a snapshot of historical data as current data was still not available in soft copy. This is because PPB was in the process of automating the process of Import / export permit issue hence going forward, data shall be obtained directly from the automated system. The study utilised all data available in electronic format on imports of pharmaceutical products into the country and only the data contained in the data entry system.

Chapter 2

LITERATURE REVIEW

2.0. Knowledge Discovery Concepts

Although there have been technological advancements in Artificial intelligence that have radically surpassed the capacity and performance of statistical methods in reporting and extraction of knowledge from databases, the level of adoption of knowledge discovery methods in Government institutions (including PPB) is wanting.

Data mining is a key phase in the process of knowledge discovery in databases that is used in the creation of models from the mass data thus producing meaningful information (Silwattananusarn & Tuamsuk, 2012). Data mining involves integration of techniques from multiple disciplines such as database and data warehousing technology, statistics, machine learning, pattern recognition, artificial neural networks, data visualization, information retrieval, image and signal processing (Wahbeh, Al-Radaideh, Al-Kabi, & Al-Shawakfa, 2011). Through the whole process of Knowledge discovery, organisations can unlock the wealth contained its data thus justifying the cost of development and implementation of expensive automated systems that may just dump data to a repository or database.

As per Figure 1, the process of knowledge discovery is described pictorially and it indicates that data mining is only a section in the whole process (Sowan, 2011). Knowledge discovery is iterative process as one may need to go back and forth in order to obtain the required knowledge from data.

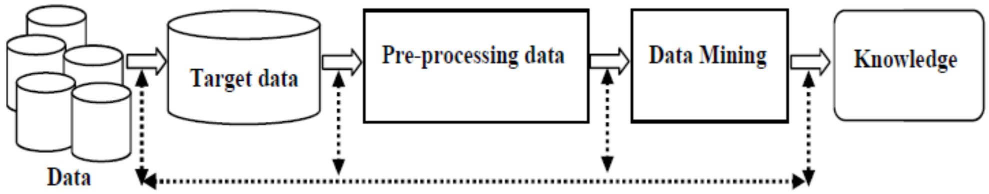


Figure 1: The Generic process of Knowledge discovery

This study covered the entire process as shown in Figure 2 so as to develop a suitable model for the discovery of knowledge for the PPB. The value of such vast amounts of data needs to be unlocked by the use of data mining in order to utilise its potential since it is not enough just to store data without using it to learn from previous occurrences. The Data mining phase can be divided into the following;

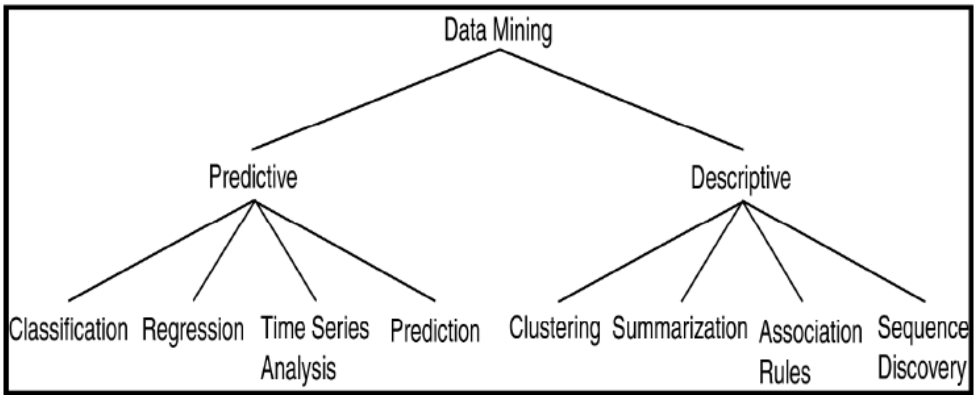


Figure 2: Broad classification of Data mining tasks within the process of Knowledge Discovery

The process of acquiring knowledge from such vast repositories of data has five major components namely Association rules, Classification or clustering, Characterization & Comparison, Sequential Pattern Analysis, Predictive analysis and Trend Analysis all of which can be embedded into powerful tools for analysis and eventual knowledge discovery (Saxena & Rajpoot, 2009). For this study focus is on descriptive data mining techniques to analyse the data of interest using correlation and association analysis.

2.1. Correlation

A correlation is a number between -1 and +1 that measures the degree of association between two attributes (call them X and Y). A positive value for the correlation implies a positive association. In this case large values of X tend to be associated with large values of Y and small values of X tend to be associated with small values of Y. A negative value for the correlation implies a negative or inverse association. In this case large values of X tend to be associated with small values of Y and vice versa. Suppose we have two attributes X and Y, with means X' and Y' respectively and standard deviations S(X) and S(Y) respectively. The correlation is computed as summation from 1 to n of the product (X(i)-X').(Y(i)-Y') and then dividing this summation by the product (n-1).S(X).S(Y) where n is total number of examples and i is the increment variable of summation.

2.2. Association

An association rule is an implication A⇒B where A and B are item sets with no items in common meaning that when A appears, B too tends to appear. The confidence of an association rule is the proportion of transactions containing A which also contain B. The discovery of relationships linking two items is desirable although for more interesting information from such a data set, it would need more than two items. (Berzal, Cubero, Marín, Sánchez, Serrano, & Vila, 2005) Association analysis is basically a Market basket analysis of the frequent item sets occurring in each instance of a transaction.

2.3. Review of Knowledge Discovery Tools

Knowledge Discovery tools exists in both commercial and open source realms. They include Oracle Data Mining, SPSS Modeller (IBM Corporation, 2012), KNIME, WEKA and Knowledge Extraction based on Evolutionary Learning (KEEL).

Knowledge discovery is a process that is invoked due to the need to gain knowledge from vast data. The data mining phase in the entire process is the most important. There are data mining problem types which are generally classified into; classification, estimation, prediction, association rules, clustering and visualization (Sharma, Osei-Bryson, & Redmond, 2008).Prediction entails the forecasting of one variable from the other known values of the other variables. Most Data mining tools are used to solve the problems although the ease of use effort and cost may vary. These tools and software provide a set of methods and algorithms that help in better utilization of data and information available to users; including methods and algorithms for data analysis, cluster analysis, Genetic algorithms, Nearest neighbour, data visualization, regression analysis, Decision trees, Predictive analytics, Text mining, etc.

The process of knowledge discovery requires the use of powerful and precise tools. Currently, the number of available tools in the market is steadily growing and thus it is becoming increasingly challenging to choose the most suitable tool for the knowledge discovery process (Mikut & Reischl, 2011).

The study proposed to use totally open source tools for the entire process of Knowledge Discovery since there is a strong trend toward this direction due to faster bug fixes and methodological improvements, potential for integration with other tools, the existence of developer and user communities, faster adoption of methods to other innovative applications, and the fair comparison of new data mining algorithms with alternative ones (Mikut & Reischl, 2011). This study reviewed some of the available knowledge discovery tools for their characteristic features and capabilities.

The current trend in the technology world is towards open source software and the adoption (Odongo, 2012). Open source software can benefit organisations by lowering initial and on-going costs and therefore allowing for flexibility. The study followed the same route in this project. This is mainly because of the trend which the Kenya Government is making as a step towards efficiency, transparency and cost reduction as proprietary tools are at times quite costly in terms of purchase price and licences. The following is a brief description of data mining tools currently available.

2.3.1. WEKA

WEKA is an acronym for Waikato Environment for Knowledge Analysis. WEKA is an open source data mining tool which is freely available from the World Wide Web under the GNU General Public License. It consists of a wide array of state of the art data mining and machine learning algorithms which are implemented in Java (Mikut & Reischl, 2011). The above algorithms can either be applied directly to a dataset or called from Java code (Hall, Frank, Holmes, Pfahringer, Reutemann, & Witten, 2012). WEKA is recognized as a landmark system in data mining and machine learning. It has achieved widespread acceptance within academia and business circles, and has become a widely used tool for data mining research.

WEKA has several tools incorporated into its structure namely: Data pre-processing, Classification, Regression, Clustering, Association rules and Visualization. With the aforementioned tools, WEKA has a detailed graphical user interface that enables easy access to its underlying functionality. It has a modular and extensible architecture allowing users to test and compare different machine learning methods on new data sets while building sophisticated data mining processes from the wide collection of base algorithms. This tool has a simple API and plug-in mechanisms that support integration of new learning algorithms with its graphical user interfaces. In addition, WEKA allows loading of data from various sources and file formats which are currently available.

2.3.2. KEEL

KEEL is an acronym for Knowledge Extraction based on Evolutionary Learning. It is an open-source software available freely on the World Wide Web which supports data management and a designer of experiments while paying special attention to the implementation of evolutionary learning and soft computing based techniques for Data Mining problems including regression, classification, clustering, and pattern mining (Alcalá-Fdez, et al., 2011).KEEL is implemented in

Java and empowers the user to analyze the behaviour of evolutionary learning for different kinds of DM problems such as regression, classification, unsupervised learning.

Evolutionary algorithms are optimization algorithms based on natural evolution and genetic processes and are considered as one of the most successful search techniques for cryptic challenges in artificial intelligence (Alcala-Fdez, et al., 2011). KEEL has the capability of performing the following types of analysis; Regression, Classification, Clustering, Pattern mining and Un-supervised learning. KEEL contains a library with evolutionary learning algorithms based on different paradigms and simplifies the integration of evolutionary learning algorithms with different pre-processing techniques thus reducing programming work. It requires less technical work thus enabling researchers to focus more on analysis of new learning models in comparison with the existing ones. The KEEL library and software were developed through the use of a strict object oriented approach thus can be used in any machine with Java and can work on any machine regardless of the operating system. Finally, it contains a user friendly interface which is oriented to the analysis of algorithms.

Screenshots of KEEL

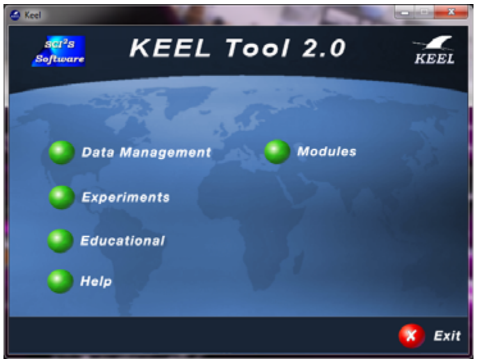
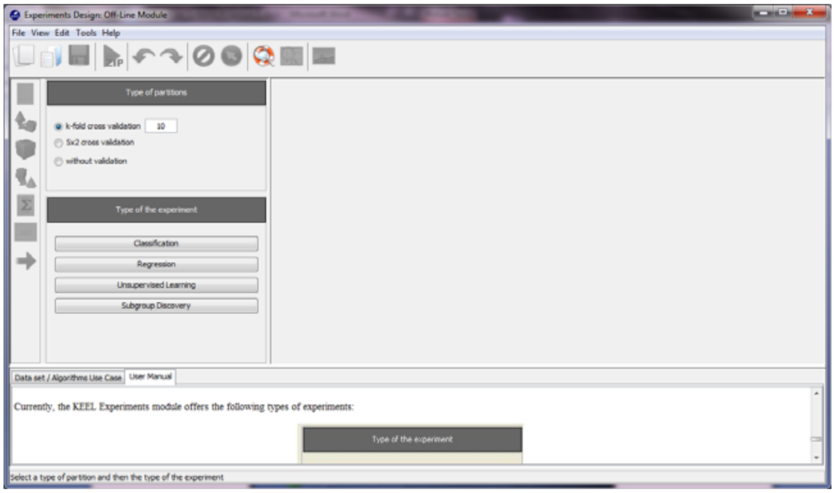


Figure 3: KEEL home screen



● k-fold cross validation 10

Currently, the KEEL Experiments module offers the following types of experiments:

Select a type of partition and then the type of the experiment

Figure 4: Working page of KEEL showing choice of experiment and a user manual at the bottom with relevant help on the specific section

2.3.3.KNIME

KNIME is an acronym for Konstanz Information Miner. It is also an open-source tool freely available in the World Wide Web. KNIME is a tool that allows a user to perform sophisticated statistics and data mining on data so as to analyze trends and predict potential results (Mikut & Reischl, 2011). It is a visual workbench that combines data access, data trans-formation, initial investigation, powerful predictive analytics and visualization (KNIME.com AG). It also provides the ability to develop reports based on one's information or automate the application of new insight back into production systems (KNIMEtech). It can perform the following types of analysis; Regression, Classification, Clustering, Pattern mining and Un-supervised learning.

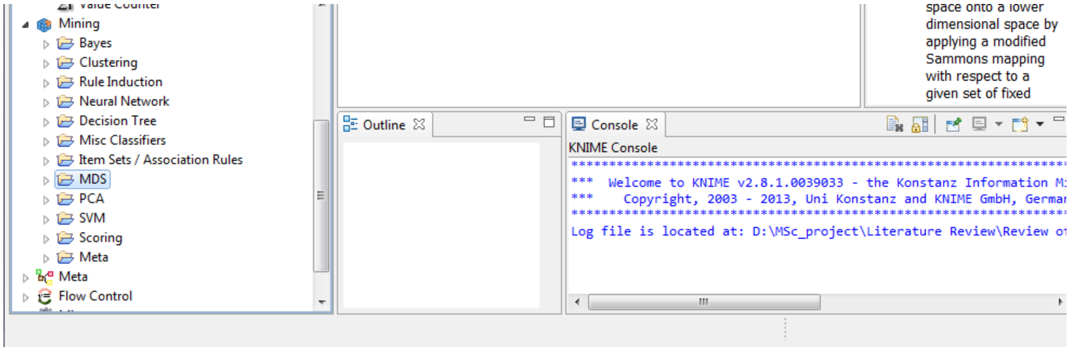
KNIME has a quite intuitive graphical user interface thus quite easy to operate. It contains an open integration platform which provides over 1000 modules including those on the KNIME community. It allows for parallel execution on multi core systems thus utilising system resources efficiently. In addition, it is highly extensible.

Screenshots of KNIME

Version 2.8.1
(Build August 21, 2013)

Copyright, 2003 - 2013, KNIME GmbH, Germany, http://www.knime.org/, contact@knime.org

Figure 5: KNIME splash screen



space onto a lower dimensional space by applying a modified Sammons mapping with respect to a given set of fixed

KNIME Console

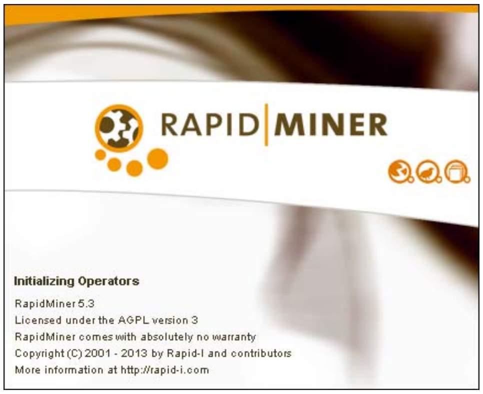
```
*****************************************************************
***  Welcome to KNIME v2.8.1.0039033 - the Konstanz Information M:
***    Copyright, 2003 - 2013, Uni Konstanz and KNIME GmbH, Germar
*****************************************************************
Log file is located at: D:\MSc_project\Literature Review\Review of
```

Figure 6: KNIME Snapshot showing its Data mining capability

2.3.4.RapidMiner

RapidMiner is a knowledge discovery tool developed in Java. It one of the world's leading open source system for Data mining that can be used as a standalone or integrated into other products. RapidMiner is a complete analytics workbench with a strong focus on data mining, text mining and predictive analytics. It provides more than 400 data mining operators thus making it able to perform numerous types of analysis. It has several modes of access; Graphical User Interface, server mode and Java API mode thus flexible. Rapidminer allows for XML process exchange and has a hands-on data mining application on the internet (online). The tool can run on major operating systems and platforms such as Windows, Linux and Macintosh with ease. It supports access to data sources like Excel, Access, Oracle, IBM DB2, Microsoft SQL, Sybase, Ingres, MySQL, Postgress, SPSS, dBase, Text files through a very simple process making it very adaptable to almost all data sources commonly used. Rapidminer has the support of "Drag and Drop" for loading of data sources and operators making it easy to learn and work with even for non–programmers. In addition, there is high availability of literature, books, websites and video tutorials on how to use it in the process of data mining as there is a large community of users of Rapidminer in the world.

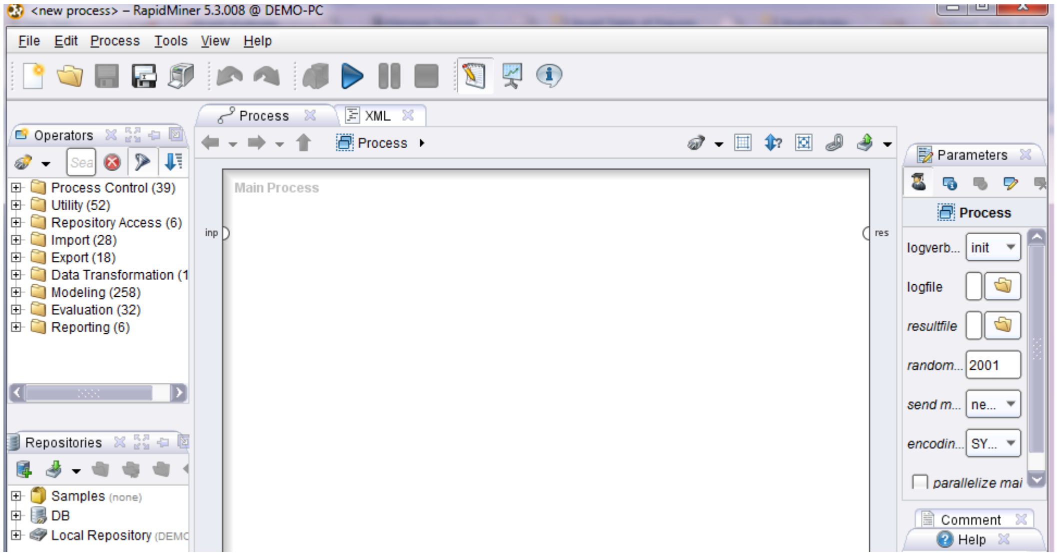
Screenshots of RapidMiner



Initializing Operators

RapidMiner 5.3
Licensed under the AGPL version 3
RapidMiner comes with absolutely no warranty
Copyright (C) 2001 - 2013 by Rapid-I and contributors
More information at http://rapid-i.com

Figure 7: Splash screen of RapidMiner currently installed



logverb... init
logfile
resultfile
random... 2001
send m... ne...
encodin... SY...

Figure 8: Working screen of RapidMiner showing its operators and tools available for use

2.4. Knowledge Discovery in the Pharmaceutical field

With the increase in population in Kenya, the data generated by the pharmaceutical industry is growing at a very high rate. These enormous and ever growing datasets present challenges and opportunities (Kusiak & Shah, 2006). The challenges include data management and knowledge discovery as traditional methods may not be applicable due to the vast nature of the datasets. The pharmaceutical industry contains very sensitive data and is one of the most data driven industry and therefore getting value from the large datasets is paramount to its success. In current times, technology is being used by pharmaceutical firms in the field of inventory management and the development of new products and services. The pharmaceutical industry mainly relies on decision oriented and systemic selection models which enable decision makers to evaluate the expected results of management decisions. A firm's competitive advantage and decision making ability can be greatly increased through the understanding of knowledge hidden in pharmaceutical data

(Ranjan, 2007).This can be done through the application knowledge discovery to data repositories held by organisations. Data mining is currently being utilised in the development of solutions such as;

i. Drug discovery technology
ii. Improved marketing strategies
iii. Decision support

Studies have also been done on the association of the various pharmaceutical product groups especially in hospitals. This has been done to determine prescription patterns of the specific medicines within the pharmaceutical product groups. Studies done in Ethiopia show that the average number of drugs per prescription ranges from 1.98 to 2.24 (Angamo, Wabe, & Raju, 2011) and this goes to show that a visit to the hospital or pharmacy outlet will give a person more than one type of drug.

Studies by (Kajungu, et al., 2012) shows that indeed there is irrational use of drugs in Tanzania which is quite similar to Kenya. The study was aimed at understanding prescribing patterns in Health facilities in health facilities in order to reduce the rate of irrational use of drugs. The study utilised classification trees which are generally easy to represent information obtained from analysis. Kajungu *et al* propose that data mining is paramount in the process of the identification and control of polypharmacy. While polypharmacy may not necessarily be wrong, the practice may place patients at high risk of adverse drug reactions and increased bacterial resistance. The same phenomenon is replicated in a study in Nigeria by (Okoro & Shekari, 2013) where there is a high occurrence of polypharmacy. The study also shows that the occurrence of an antibiotic in a prescription to a patient is quite high at 56.2%.

According to a study by (Viktil, Blix, Moger, & Reikvam, 2006) polypharmacy can be linked to Drug Related Problems (DRP). The study confirms that there is a linear relation between the numbers of drugs prescribed per admission and Drug Related Problems. An increase in the number of drugs causes a corresponding increase in Drug Related Problems. This fact makes it critical for researchers to look into the analysis of prescription patterns of most pharmaceutical practitioners so as to help protect patients from adverse effects of drug reactions.

Numerous studies on prescription data mining show that information obtained is quite relevant to the concerned regulatory authorities to formulate the required policy to safeguard the health of the citizens under their jurisdiction. This study proposed to perform analysis of imports of pharmaceutical products in order to investigate whether the patterns observed from the prescriptions actually exist within data on imports.

Imports of Pharmaceutical products are assumed to satisfy the demand for the various institutions dispensing them and thus the imports can be linked to the needs of the country at a high level.
Data mining of medicine information at the Pharmacy and Poisons Board can be vital in the process of identification of patterns of movement of particular medicines, the anomalies in the distribution chain and deficits or malpractices which need regulatory action from the disciplinary arm of the Board. The study proposes to provide the Board with a model which can be of great value in the detection of anomalies and malpractices in the distribution chain of controlled medicines through the detailed analysis of historical data.

Domain-Driven Data Mining (D 3M) aims to develop general principles, methodologies, and techniques for modelling and merging comprehensive domain-related factors and synthesized ubiquitous intelligence surrounding problem domains with the data mining process, and discovering knowledge to support business decision-making (Silwattananusarn & Tuamsuk, 2012). There are several algorithms that can be used in the field of data mining depending on the type of data under analysis. Vast amounts of data held by PPB can only be useful only when patterns are discovered from it thus the nature of the problem domain can be revealed.

The patterns can then be utilised to solve specific problems as being interpreted or inferred with (Wang & K. C. Wong, 2003). Despite the fact that there are advancements in computer technologies, the process of extraction of information and knowledge contained in data from organisations is a difficult problem. Many systems produce lots of data which at times is poor in information (Wang & K. C. Wong, 2003). The main issue is actually finding the needle in the hay stack. Due to the above, the best approach is to thin out the search domain so as to identify the classification rules inherent in the data. To handle this, it is necessary to perform pre-processing, filtering and attribute reduction so as to reduce noisy data and bring out useful information from the data.

2.5. Conceptual Model

The Knowledge discovery process could also be described by the model shown in figure 9.

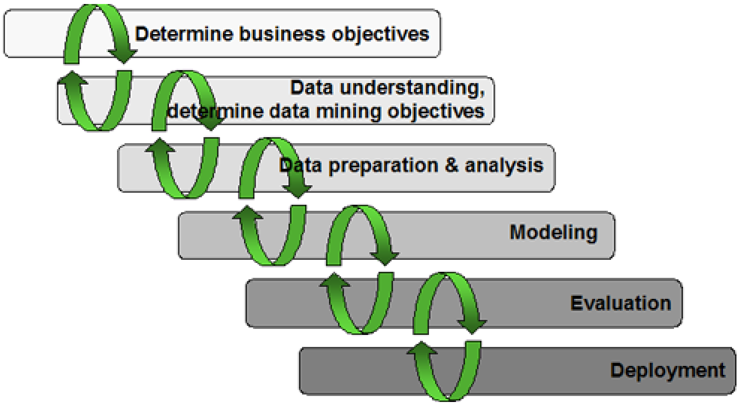

Determine business objectives
Data understanding, determine data mining objectives
Data preparation & analysis
Modeling
Evaluation
Deployment

Figure 9: Process steps of Knowledge discovery

Source: http://www.microsegment.hu/ms/methodology1.php

As seen from the above models, it is clear that the knowledge discovery process is quite iterative in nature since a researcher may need to refer back to previous sections after receiving output from advanced sections. The steps are outlined below.

2.5.1. Business Understanding

This initial phase focuses on understanding the project objectives and requirements from a Business perspective and then converting this knowledge into a data mining problem definition, and a preliminary project plan designed to achieve the objectives.

2.5.2. Data Understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data. Its aim is to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information. There is a close link between Business Understanding and Data Understanding. The formulation of the data mining problem and the project plan require at least some understanding of the available data.

2.5.3. Data Preparation

Data preparation is very important in the process of data mining. The data preparation phase covers all activities to construct the final dataset (data to be loaded into the modeling tool(s)) from the initial raw data. Raw data is usually noisy, incomplete or impure which is very likely to hamper the discovery useful patterns Current data mining tools require high quality data and high quality data results in intelligent patterns (Zhang, Zhang, & Yang, 2003). The study intends to obtain a dataset that is quite refined and accurate for the process of data mining so as to obtain meaningful results which can help management to make informed decisions from the vast amounts of data they hold. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection, data cleaning, construction of new attributes, and transformation of data for modelling tools.

2.5.4. Modelling

In this phase, various modelling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type as some techniques require specific data formats. There is a close link between Data Preparation and Modelling. Often, one realizes data problems while modelling or one gets ideas for constructing new data.

2.5.5. Evaluation

At this stage evaluation of models built are thoroughly evaluated in order to ascertain that it properly answers the business questions and satisfies the objectives set out at the beginning. Once this is complete, the model can now be used by the customer to solve the problem or satisfy the objectives set out.

2.5.6. Deployment

The process of creation of data mining models is not the end of the whole exercise of data mining. After the creation of the model, the usefulness of the model needs to be utilised by the customer (Džeroski, 2007), thereby justifying why so much time and resources have been spent on the project. The aim of this phase is to hand over the model to the customer in a way that he/she can utilise it on other similar scenarios with the same data source.

Chapter 3

METHODOLOGY

3.0 Introduction

This section begins by providing information on the tools that were used in the study. It is then followed by a description of how the research was done based on the CRISP-DM model of data mining.

3.1 Research design

The study utilised quantitative research. RapidMiner and Microsoft SQL Server were used to analyse the PPB database containing data on imports and exports of pharmaceutical products to understand and give out intelligent inferences from it. The above mentioned tools were utilised according to the CRISP-DM model for carrying out similar projects.

3.2 Data Sources

The study utilised data held by the PPB import export database. This data was held by the Licit control Department of the Board that is charged with the issuance of import / export permits regarding pharmaceutical products in Kenya. This database contains information pertaining to quantities of imports and exports products and the respective dates of permits issued. In addition, it also contains the business entities to which the permits have been given.

3.3 Tools

According to literature review, the study utilised RapidMiner as the tool for this project. It was downloaded and installed on a personal computer external to PPB. This tool was used to connect to Microsoft SQL Server where queries were loaded onto the tool thus feeding it with the relevant data to produce results of the project.

3.4 Methods

The CRISP-DM Methodology

The CRISP-DM (CRoss Industry Standard Process for Data Mining) project proposed a comprehensive process model for carrying out data mining projects. The process model is independent of both the industry sector and the technology used (Wirth & Hipp, 2000). It is one of the most popular and broadly adopted model for knowledge discovery. It has already been acknowledged and relatively widely used especially in the fields of research and industrial communities (Kurgan & Musilek, 2006). The aim of CRISP-DM is to provide an efficient process that can be used by persons with lower technical skills in data mining to produce knowledge from their vast data repositories.

CRISP-DM begins by carrying out an analysis of a business problem for transforming it into a technical data mining problem. CRISP-DM can also be integrated with a specific project management methodology complementing administrative and technical tasks. It is also widely distributed at no cost, unlike SEMMA (SAS,2009b). CRISP-DM defines a structure for data mining projects and provides orientation for their execution. It serves both as a reference model and a user guide (Chapman et al., 2000). The reference model gives a general view of a data mining project's life-cycle, containing each phase with its objective, the tasks, the relationships between them and the step-by-step instructions that must be carried out. It is proposed that the study will utilise the CRISP-DM standard as it will guide the whole process from start to finish. Figure 10 outlines the basic phases of the CRISP-DM methodology. The outer circle in Figure 10 is indicative of the fact that the process of knowledge discovery is not complete once the output solution is deployed but the lessons learnt from the solution may be put to use in subsequent knowledge discovery exercises. The following section will outline how the study will be done as per phase of the model.

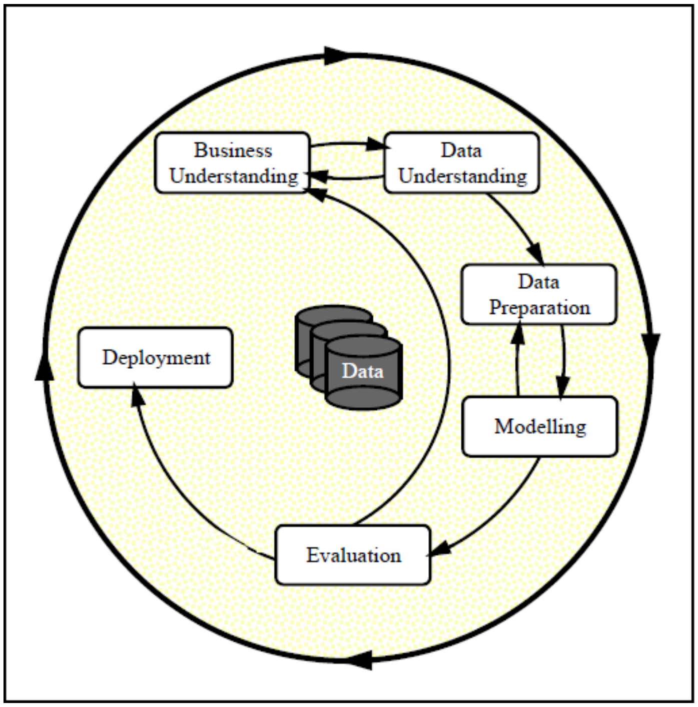


Figure 10: Phases of the CRISP-DM process model for the entire knowledge discovery process

3.4.1. Business Understanding Phase

This phase was the key linking factor of this project to the Project site in that it relates the Project objectives and requirements to the business perspective of the organisation. The knowledge obtained is then converted into a data mining problem definition. The study will identify key persons in the import and export section, identify other sections that are likely to be impacted by the proposed project, gather user requirements and expectations thereby describing the project problem in general. The study looked into whether PPB were using any data mining, the hardware and software present relating to data on controlled medicines and the relevant background of the situation at hand.

This phase was to be re-visited since the CRISP-DM model allows for repeated inference of any phase since as the process continues, some issues may arise that may need the researcher to get a better understanding of the business. In this phase we will relate the business questions into data mining goals while specifying the data mining problem type and the criteria for model assessment. Examples of data mining problem types include classification, prediction, association and clustering. The business process for the issuance of import and / or export permits to agents is as shown in Figure 11

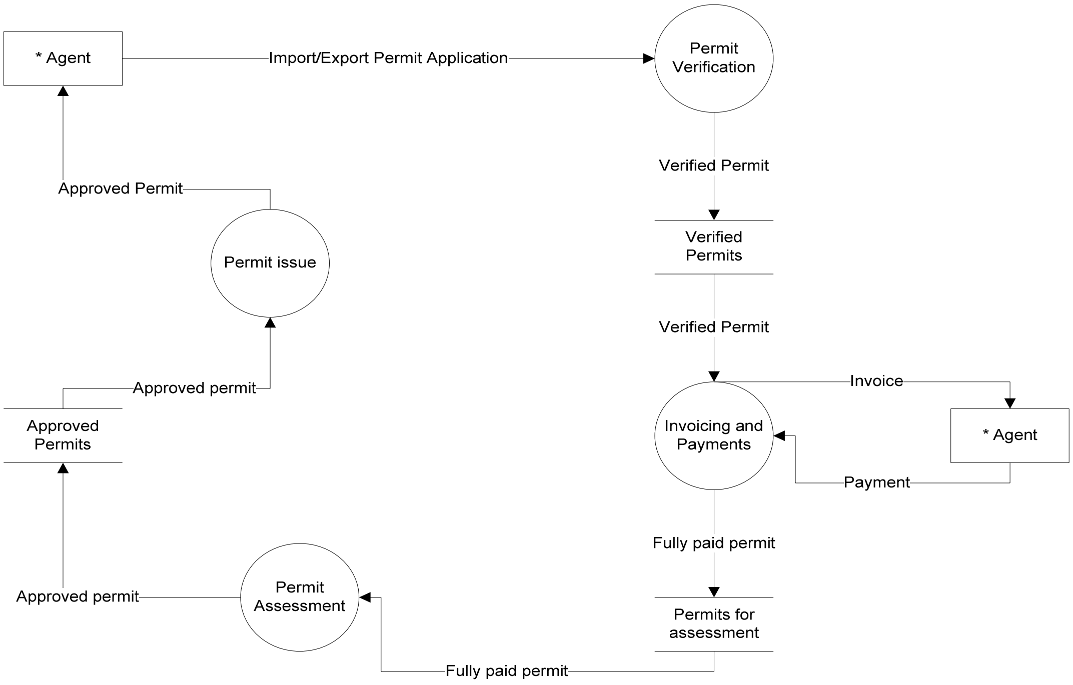


Figure 11: Data flow diagram representing the process of the application of a permit till issue

An agent applies for a permit based on their need and the application is accompanied by a proforma invoice from the exporter if the transaction is import and an Import Declaration Form

(IDF) from the Kenya Revenue Authority (KRA). The IDF and proforma invoice provide information on the cost of the consignment down to the constituent items and this is the value the study is using. The permit applications are then verified for completeness and then the agent is advised to pay the fee for the permit which is usually 0.75% of the IDF value after which the permit application is assessed and approved or rejected according to PPB policy and guidelines. On approval, the agent of the respective application is notified and is issued with the permit and can now go to the particular port of entry to clear the consignment. A repository of copies of all issued permits is retained at PPB for future reference. In case of rejection, the permit application is returned to the agent to rectify or clarify any details that may be unclear and then the permit may follow the same procedure till approval. In some cases, permit applications may be rejected and the agent may be required to start a fresh application. The data of concern to the study was only of those permits that were approved. Of major concern to PPB was on obtaining valuable information from the data on approved permits. It was important to

i. Obtain yearly market trends of the various categories of pharmaceutical products
ii. Discover associations if any within the pharmaceutical product categories with time

3.4.2. Data Understanding Phase

Data on imports and exports permits was available from the electronic data repository from the data entry system. The database is housed in Microsoft Access and contains several tables and the tables have several attributes. The following are the tables as they appear in the repository of the data entry system. Table 1 represents the tables present.

Table 1: Summary of the PPB import / export database

Table name	Information contained
dbo_dbo_Agent	Stores basic information of the agent such as City, Postal Address, Telephone Number and email
dbo_dbo_invoice	Stores information on the Proforma Invoice of each consignment, the manufacturer, the agent importing the consignment, the origin country and port, destination country and port, and the Invoice Value of the consignment
dbo_dbo_Manufacturer	Stores basic information of the Manufacturer such as City, Postal Address, Telephone Number and email
dbo_dbo_Country	Stores a list of all countries in the world
dbo_dbo_LineItem	Stores detailed information of the products imported/exported within each proforma Invoice. It is linked to dbo_dbo_invoice via a foreign key and is also linked to dbo_dbo_IDF via foreign key
dbo_dbo_IDF	Stores data from the IDF details such as the IDF total value, IDF number and the date of import

Figure 12 shows an illustration of the various relationships that exist in the source database that were useful in the generation of queries to the database.

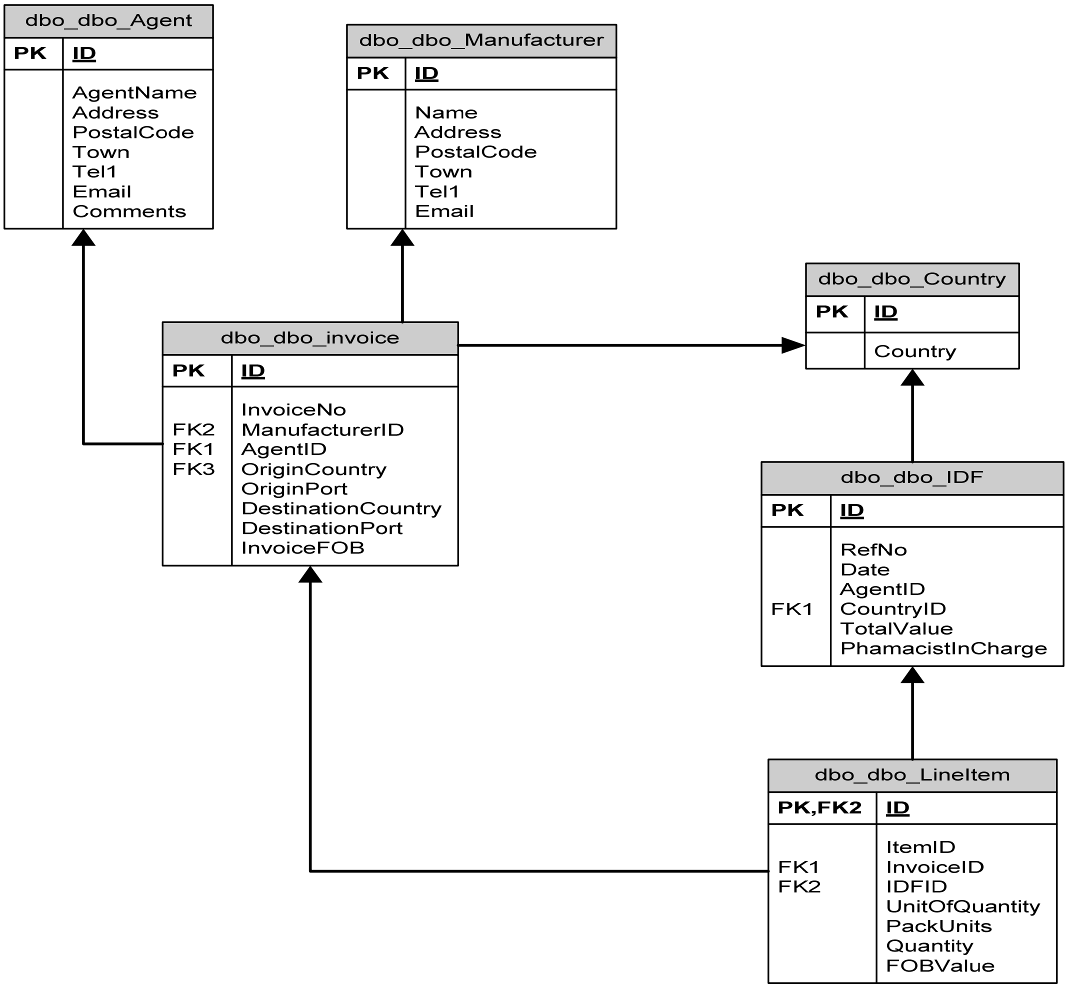


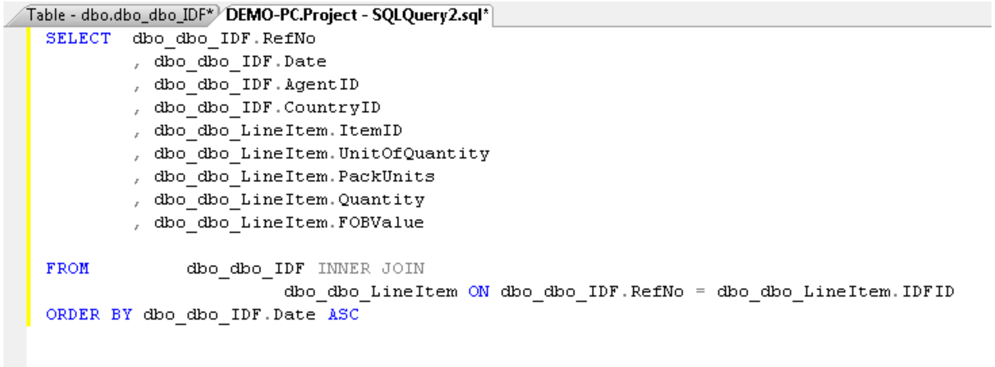
Figure 12: Entity relationship diagram of the PPB import / export database

3.4.3.Data Preparation

A full backup of the database was taken from PPB containing all the tables and migrated to Microsoft SQL Server through the Upsizing wizard present in Microsoft Access. Based on the relationships outlined in the data understanding phase, specific database scripts were developed to query the database of relevant data for the study. The database scripts contained a number of joins to produce data useful for the analysis.

Issues found

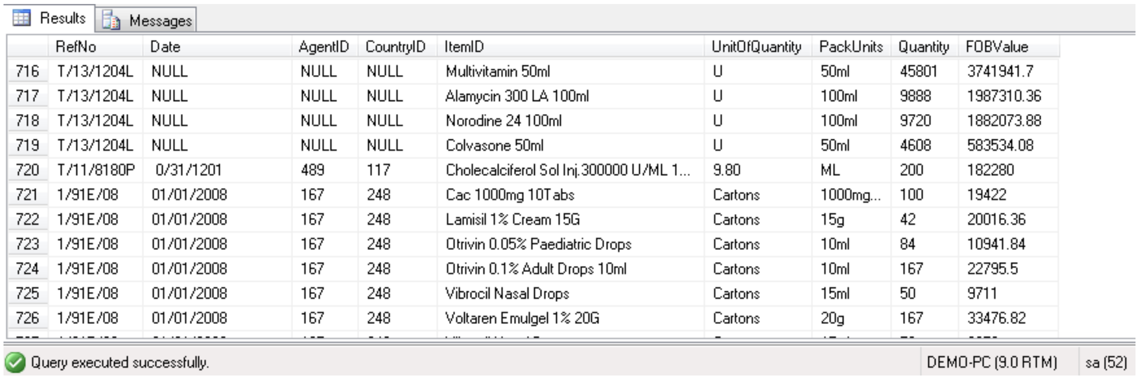
On running the following database script as shown in Figure 13, several issues were discovered such as missing values and inconsistent dates as shown in figure 14.



```
SELECT  dbo_dbo_IDF.RefNo
        , dbo_dbo_IDF.Date
        , dbo_dbo_IDF.AgentID
        , dbo_dbo_IDF.CountryID
        , dbo_dbo_LineItem.ItemID
        , dbo_dbo_LineItem.UnitOfQuantity
        , dbo_dbo_LineItem.PackUnits
        , dbo_dbo_LineItem.Quantity
        , dbo_dbo_LineItem.FOBValue

FROM         dbo_dbo_IDF INNER JOIN
                      dbo_dbo_LineItem ON dbo_dbo_IDF.RefNo = dbo_dbo_LineItem.IDFID
ORDER BY dbo_dbo_IDF.Date ASC
```

Figure 13: Database script for querying the dataset for data mining



	RefNo	Date	AgentID	CountryID	ItemID	UnitOfQuantity	PackUnits	Quantity	FOBValue
716	T/13/1204L	NULL	NULL	NULL	Multivitamin 50ml	U	50ml	45801	3741941.7
717	T/13/1204L	NULL	NULL	NULL	Alamycin 300 LA 100ml	U	100ml	9888	1987310.36
718	T/13/1204L	NULL	NULL	NULL	Norodine 24 100ml	U	100ml	9720	1882073.88
719	T/13/1204L	NULL	NULL	NULL	Colvasone 50ml	U	50ml	4608	583534.08
720	T/11/8180P	0/31/1201	489	117	Cholecalciferol Sol Inj.300000 U/ML 1...	9.80	ML	200	182280
721	1/91E/08	01/01/2008	167	248	Cac 1000mg 10Tabs	Cartons	1000mg...	100	19422
722	1/91E/08	01/01/2008	167	248	Lamisil 1% Cream 15G	Cartons	15g	42	20016.36
723	1/91E/08	01/01/2008	167	248	Otrivin 0.05% Paediatric Drops	Cartons	10ml	84	10941.84
724	1/91E/08	01/01/2008	167	248	Otrivin 0.1% Adult Drops 10ml	Cartons	10ml	167	22795.5
725	1/91E/08	01/01/2008	167	248	Vibrocil Nasal Drops	Cartons	15ml	50	9711
726	1/91E/08	01/01/2008	167	248	Voltaren Emulgel 1% 20G	Cartons	20g	167	33476.82

Query executed successfully. DEMO-PC (9.0 RTM) sa (52)

Figure 14: Screenshot of returned data showing issues that need to be resolved before data mining activities

Missing values were represented as 'NULL' in the respective rows where they were found. There were missing values for 'FOBValue', 'Quantity', ItemID', 'Date', 'RefNo', 'CountryID' and 'AgentID'. This issue presented data that could not be linked to any of the mentioned attributes.

Inconsistent and erroneous date values for example '03-12-2011' and '04/01/2008' indicating different ways of representing date while '04/02/2023' indicating futuristic dates. This issue made it not possible to segment the data into years as the dates were captured as data type 'CHAR' and not 'DATETIME'. Classification of the various pharmaceutical products into predefined groups was not done and therefore for ease of analysis as per the requirements from PPB this was necessary.

Solutions and Transformations

Missing Values

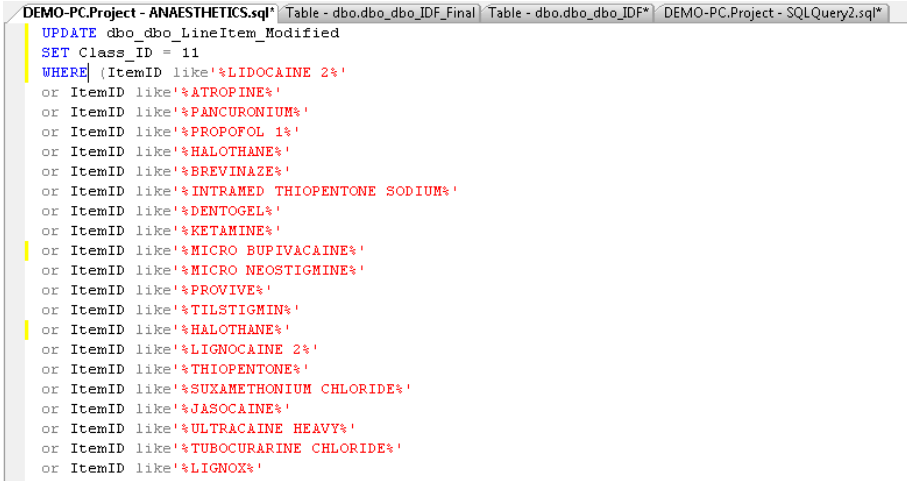
Upon discussion with domain experts on the proposed analysis to be done, it was agreed to omit all values where 'ItemID' is NULL, 'CountryID' is NULL and 'FOBValue' is NULL

Inconsistent Date values

The table ‘dbo_dbo_IDF’ was exported to Microsoft Excel where by the column containing the date was formatted to only contain ‘Date’ values. Once this was done, then the excel sheet was saved and imported back to MSSQL Server ‘dbo_dbo_IDF_Final’

Lack of Classification of Pharmaceutical Products

Classification is usually done by the domain experts who are usually pharmacists and pharmaceutical technologists who are conversant with the Kenyan market. They usually match the names on the product to a list of existing group for example in the group ‘Antibiotic’ there are products having the name or part of their name containing ‘Amoxil’, a product having part of its name as ‘Malarid’ is likely to fall within the category if ‘Anti-Malarials’ and so on. Due to the large number of records in the database, it was close to impossible to classify the items one by one through a pharmaceutical expert. With this knowledge of the technique of matching, a sample of all product names and their respective group names was obtained from PPB and this was used to generate a database script that was used to group all the products into their respective classifications. A new table was created to contain the names of all product categories and their category ID. In order to maintain the original table ‘dbo_dbo_LineItem’, a new table was created called ‘dbo_dbo_Lineitem_Modified’ which had a new column for the ‘Class_ID’. A sample script used is shown in Figure 15.


DEMO-PC.Project - ANAESTHETICS.sql*
Table - dbo.dbo_dbo_IDF_Final
Table - dbo.dbo_dbo_IDF*
DEMO-PC.Project - SQLQuery2.sql*

```
UPDATE dbo_dbo_LineItem_Modified
SET Class_ID = 11
WHERE (ItemID like'%LIDOCAINE 2%'
or ItemID like'%ATROPINE%'
or ItemID like'%PANCURONIUM%'
or ItemID like'%PROPOFOL 1%'
or ItemID like'%HALOTHANE%'
or ItemID like'%BREVINAZE%'
or ItemID like'%INTRAMED THIOPENTONE SODIUM%'
or ItemID like'%DENTOGEL%'
or ItemID like'%KETAMINE%'
or ItemID like'%MICRO BUPIVACAINE%'
or ItemID like'%MICRO NEOSTIGMINE%'
or ItemID like'%PROVIVE%'
or ItemID like'%TILSTIGMIN%'
or ItemID like'%HALOTHANE%'
or ItemID like'%LIGNOCAINE 2%'
or ItemID like'%THIOPENTONE%'
or ItemID like'%SUXAMETHONIUM CHLORIDE%'
or ItemID like'%JASOCAINE%'
or ItemID like'%ULTRACAINE HEAVY%'
or ItemID like'%TUBOCURARINE CHLORIDE%'
or ItemID like'%LIGNOX%'
```

Figure 15: Database script for classification of all products that are in the category 'Anaesthetics'

The script was quite successful as it was able to classify all of the products which had ‘ItemID’ NOT NULL and therefore further analysis and modelling was possible.

Missing Data

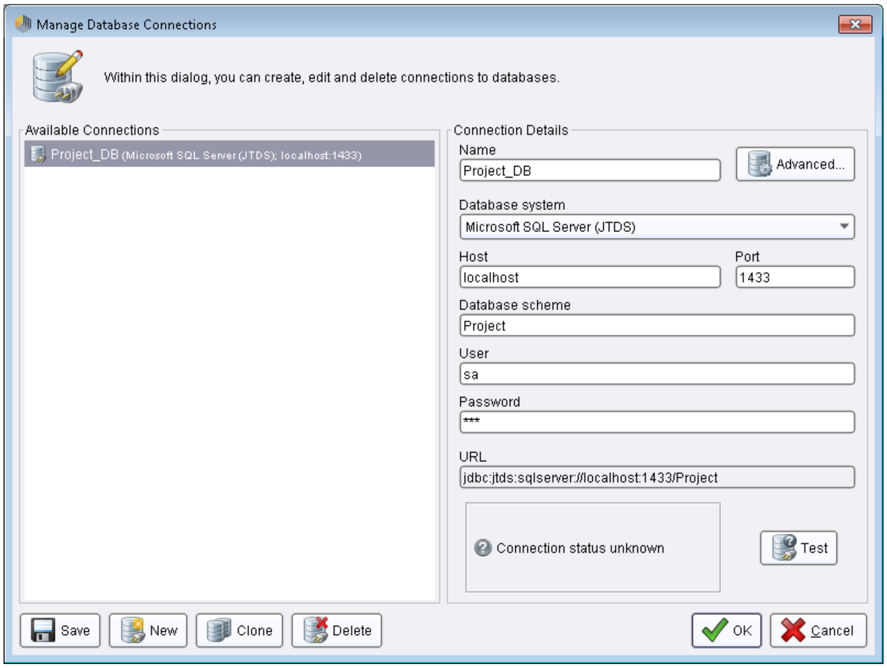
Upon modification of the date attribute of the entity ‘dbo_dbo_IDF’ it was discovered that the process of data entry skipped the years 2009 and 2010 as when the data is grouped by years, there are no records for 2009 and very few for 2010.

3.4.4. Modelling Phase

In this section, the data obtained from the Data preparation phase was fed into the RapidMiner for analysis.

Initialization of RapidMiner for analysis

To start on the process of modelling, the RapidMiner requires data. Since our data was in MSSQL Server, then the first step was to create a database connection to be used to connect RapidMiner to MSSQL Server without having to specify the details over and over again each time a new query is to be run. This was done at the interface in Figure 16



Manage Database Connections

Within this dialog, you can create, edit and delete connections to databases.

Available Connections

Project_DB (Microsoft SQL Server (JTDS); localhost:1433)

Connection Details

Name: Project_DB

Database system: Microsoft SQL Server (JTDS)

Host: localhost

Port: 1433

Database scheme: Project

User: sa

Password: ***

URL: jdbc:jtds:sqlserver://localhost:1433/Project

Connection status unknown

Figure 16: RapidMiner interface for managing database connections

The next stage is the opening of a new ‘Process’ which shall open a new working space.

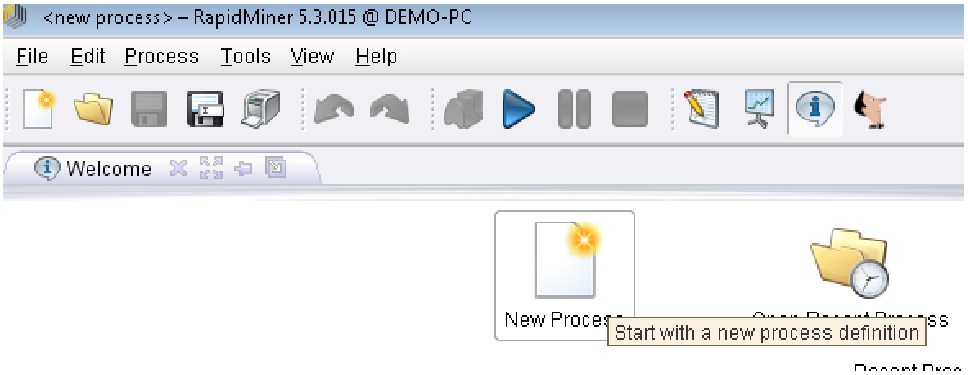
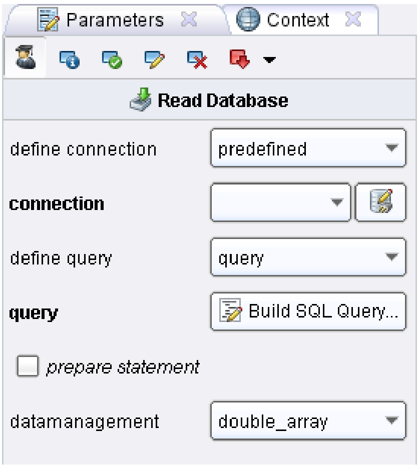

<new process> – RapidMiner 5.3.015 @ DEMO-PC
File
Edit
Process
Tools
View
Help
Welcome
New Process
Start with a new process definition
Open Recent Process

Figure 17: How to get a new workspace

Once this is done, the next step is to load data for modelling to the process. This is done with the help of ‘Import’ operators that are on the left hand side of the window. In import, there are five sub folders, but for now we need to expand the data sub folder to find an operator called ‘Read Database’. This operator is then dragged on to the workspace and easily configured on the left hand side so as it can give RapidMiner the data required.



define connection	predefined
connection	
define query	query
query	Build SQL Query...
□ *prepare statement*	
datamanagement	double_array

Figure 18: Database connection configuration

The connection configured earlier is chosen and then the query is pasted in the ‘Build SQL Query’ window that pops up when clicked. Once complete, the ‘out’ port of the operator is connected to the ‘res’ port through dragging with the mouse and when the play button is clicked to run the process, it brings up results of the query.

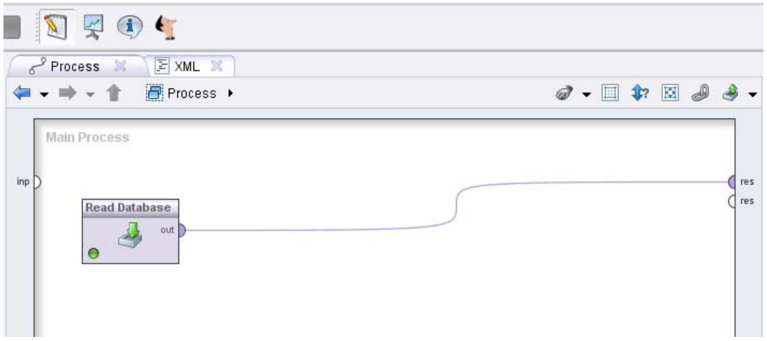
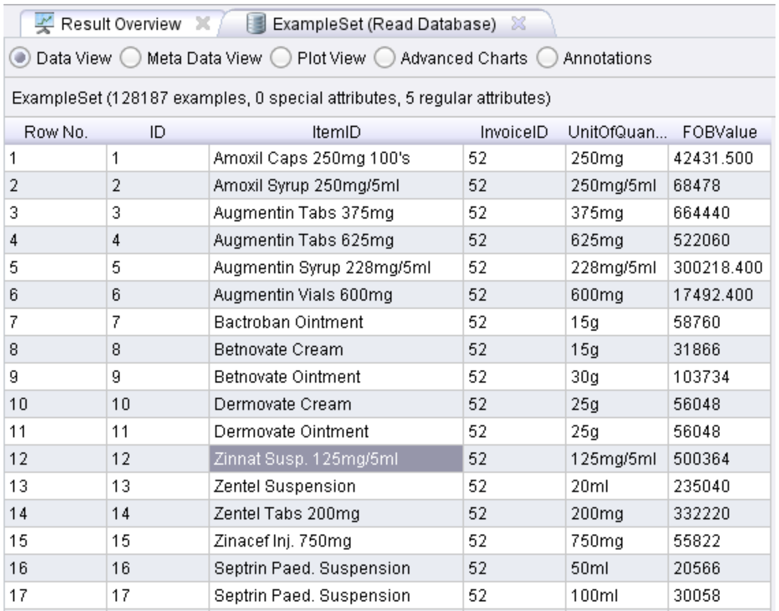


Figure 19: RapidMiner Process before execution sowing the process view



● Data View ○ Meta Data View ○ Plot View ○ Advanced Charts ○ Annotations

ExampleSet (128187 examples, 0 special attributes, 5 regular attributes)

Row No.	ID	ItemID	InvoiceID	UnitOfQuan...	FOBValue
1	1	Amoxil Caps 250mg 100's	52	250mg	42431.500
2	2	Amoxil Syrup 250mg/5ml	52	250mg/5ml	68478
3	3	Augmentin Tabs 375mg	52	375mg	664440
4	4	Augmentin Tabs 625mg	52	625mg	522060
5	5	Augmentin Syrup 228mg/5ml	52	228mg/5ml	300218.400
6	6	Augmentin Vials 600mg	52	600mg	17492.400
7	7	Bactroban Ointment	52	15g	58760
8	8	Betnovate Cream	52	15g	31866
9	9	Betnovate Ointment	52	30g	103734
10	10	Dermovate Cream	52	25g	56048
11	11	Dermovate Ointment	52	25g	56048
12	12	Zinnat Susp. 125mg/5ml	52	125mg/5ml	500364
13	13	Zentel Suspension	52	20ml	235040
14	14	Zentel Tabs 200mg	52	200mg	332220
15	15	Zinacef Inj. 750mg	52	750mg	55822
16	16	Septrin Paed. Suspension	52	50ml	20566
17	17	Septrin Paed. Suspension	52	100ml	30058

Figure 20: Results View of RapidMiner showing results similar to what MSSQL server would show

Important Buttons for changing views

▶ Used to execute a process

Used to view the process

Used to view results of a process

Initial analysis

For analysis, the process required to be loaded with data. Once loaded other required operators were dragged onto the workspace and the output connected to the ‘res’ port so as to obtain results from the joined operators. For analysis of data, the following query was prepared and loaded into the RapidMiner query window. The dates were changed appropriately so that the results could be obtained for the various years within the dataset of PPB. The Process was as shown in Figure 21.

```
SELECT  dbo_dbo_IDF_New.Date
       , SUM(CASE WHEN DATA.Class_ID = 1 THEN DATA.FOBValue ELSE NULL END) as DRUGS_AFFECTING_BLOOD
       , SUM(CASE WHEN DATA.Class_ID = 2 THEN DATA.FOBValue ELSE NULL END) as MUSCLE_RELAXANTS
       , SUM(CASE WHEN DATA.Class_ID = 3 THEN DATA.FOBValue ELSE NULL END) as ANTI_DIABETICS
       , SUM(CASE WHEN DATA.Class_ID = 4 THEN DATA.FOBValue ELSE NULL END) as BLOOD_PRODUCTS_BLOOD_SUBSTITUTES
       , SUM(CASE WHEN DATA.Class_ID = 5 THEN DATA.FOBValue ELSE NULL END) as
HORMONES_ENDOCRINE_and_CONTRACEPTIVES
       , SUM(CASE WHEN DATA.Class_ID = 6 THEN DATA.FOBValue ELSE NULL END) as RESPIRATORY_TRACT
       , SUM(CASE WHEN DATA.Class_ID = 7 THEN DATA.FOBValue ELSE NULL END) as ANTI_ALLERGICS_ANAPHYLAXIS
       , SUM(CASE WHEN DATA.Class_ID = 8 THEN DATA.FOBValue ELSE NULL END) as PSYCHOTHERAPEUTIC_DRUGS
       , SUM(CASE WHEN DATA.Class_ID = 9 THEN DATA.FOBValue ELSE NULL END) as VETERINARY_PRODUCTS
       , SUM(CASE WHEN DATA.Class_ID = 10 THEN DATA.FOBValue ELSE NULL END) as MISCELLANEOUS
       , SUM(CASE WHEN DATA.Class_ID = 11 THEN DATA.FOBValue ELSE NULL END) as ANAESTHETICS
       , SUM(CASE WHEN DATA.Class_ID = 12 THEN DATA.FOBValue ELSE NULL END) as VITAMIN_and_MINERALS
       , SUM(CASE WHEN DATA.Class_ID = 13 THEN DATA.FOBValue ELSE NULL END) as ANTIEPILEPTICS
       , SUM(CASE WHEN DATA.Class_ID = 14 THEN DATA.FOBValue ELSE NULL END) as
SOLUTIONS_WATER_ELECTROLYTE_ACID_BASE_DISTURBANCE
       , SUM(CASE WHEN DATA.Class_ID = 15 THEN DATA.FOBValue ELSE NULL END) as GASTROINTESTINAL
       , SUM(CASE WHEN DATA.Class_ID = 16 THEN DATA.FOBValue ELSE NULL END) as IMMUNOLOGICALS_VACCINES
       , SUM(CASE WHEN DATA.Class_ID = 17 THEN DATA.FOBValue ELSE NULL END) as ANTI_VIRALS
       , SUM(CASE WHEN DATA.Class_ID = 18 THEN DATA.FOBValue ELSE NULL END) as ANALGESICS_ANTIPYRETICS_NSAIDs
       , SUM(CASE WHEN DATA.Class_ID = 19 THEN DATA.FOBValue ELSE NULL END) as ANTIMIGRANE
       , SUM(CASE WHEN DATA.Class_ID = 20 THEN DATA.FOBValue ELSE NULL END) as ANTI_RETROVIRALS
       , SUM(CASE WHEN DATA.Class_ID = 21 THEN DATA.FOBValue ELSE NULL END) as
ANTINEOPLASTICS_and_IMMUNOSUPPRESSIVE
       , SUM(CASE WHEN DATA.Class_ID = 22 THEN DATA.FOBValue ELSE NULL END) as ANTI_MALARIALS
       , SUM(CASE WHEN DATA.Class_ID = 23 THEN DATA.FOBValue ELSE NULL END) as OPHTALMOLOGICAL_ENT_PREPARATIONS
       , SUM(CASE WHEN DATA.Class_ID = 24 THEN DATA.FOBValue ELSE NULL END) as DISINFECTANTS_ANTISEPTICS
       , SUM(CASE WHEN DATA.Class_ID = 25 THEN DATA.FOBValue ELSE NULL END) as DIAGNOSTICS_AGENTS_RADIOLOGICALS
       , SUM(CASE WHEN DATA.Class_ID = 26 THEN DATA.FOBValue ELSE NULL END) as ANTI_TBS
       , SUM(CASE WHEN DATA.Class_ID = 27 THEN DATA.FOBValue ELSE NULL END) as ANTIPARKINSONISM
       , SUM(CASE WHEN DATA.Class_ID = 28 THEN DATA.FOBValue ELSE NULL END) as DIURETICS
       , SUM(CASE WHEN DATA.Class_ID = 29 THEN DATA.FOBValue ELSE NULL END) as OXYTOCICS_ANTIOXYTOCICS
       , SUM(CASE WHEN DATA.Class_ID = 30 THEN DATA.FOBValue ELSE NULL END) as ANTI_BIOTICS
       , SUM(CASE WHEN DATA.Class_ID = 31 THEN DATA.FOBValue ELSE NULL END) as ANTI_FUNGALS
       , SUM(CASE WHEN DATA.Class_ID = 32 THEN DATA.FOBValue ELSE NULL END) as CARDIOVASCULAR
       , SUM(CASE WHEN DATA.Class_ID = 33 THEN DATA.FOBValue ELSE NULL END) as PERITONEAL_DIALYSIS_SOLUTIONS
       , SUM(CASE WHEN DATA.Class_ID = 34 THEN DATA.FOBValue ELSE NULL END) as DERMATOLOGICAL
       , SUM(CASE WHEN DATA.Class_ID = 35 THEN DATA.FOBValue ELSE NULL END) as
ANTIDOTES_and_SUBSTANCES_USED_IN_POISONINGS


FROM    dbo_dbo_LineItem_Modified AS  DATA INNER JOIN
                         dbo_dbo_IDF_New ON DATA.IDFID = dbo_dbo_IDF_New.RefNo INNER JOIN
                         dbo_dbo_invoice ON DATA.InvoiceID = dbo_dbo_invoice.ID INNER JOIN
                         dbo_dbo_Country ON dbo_dbo_invoice.OriginCountry = dbo_dbo_Country.ID INNER JOIN
                                            dbo_dbo_Agent on dbo_dbo_Agent.ID = dbo_dbo_IDF_New.AgentID
WHERE  dbo_dbo_invoice.Is_import = 1
        AND dbo_dbo_IDF_New.Date is not null
        AND DATA.FOBValue > 0
        AND DATA.Class_ID  IN (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22,
23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35)
        AND dbo_dbo_IDF_New.Date BETWEEN '01-JAN-2008' AND '31-DEC-2013'
GROUP BY dbo_dbo_IDF_New.Date


order by 1
```

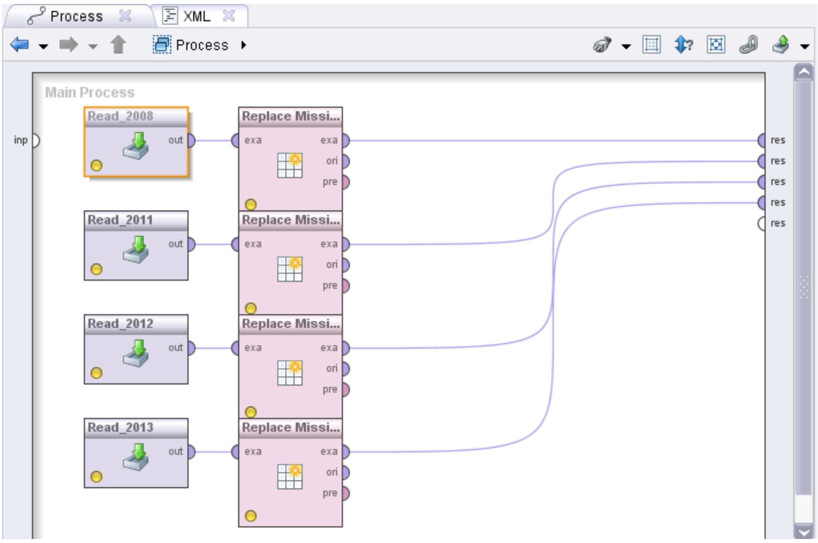


Figure 21: RapidMiner process for displaying the general information in the PPB dataset on Imports

An operator 'Replace missing values' was added to the process so as to replace the NULL values with 'zero' for ease of plotting

a. Correlation Analysis

This analysis was done on the dataset from the aforementioned query. A new process was created with the 'Read Database' operator and the 'Correlation Matrix' operator as illustrated in figure 22. The dataset range was modified to include the years 2008 through to 2013.

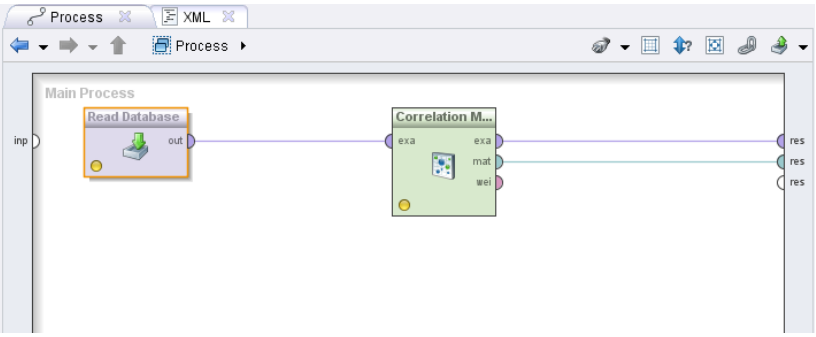


Figure 22: Correlation process in RapidMiner

b. Association Analysis

The same data set was then subjected to a new process for association analysis. This process required the following operators illustrated in figure 23.

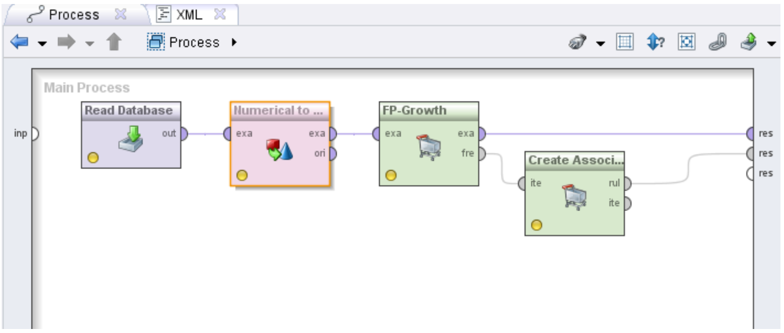


Figure 23: Process workflow for Association Analysis

i. **Read Database**

Similar to the previous ‘Read Database’ the only difference being that the query ‘WHERE’ clause has been slightly modified

ii. **Numerical to Binominal**

This operator performs data preparation on the fly as it converts any number greater than zero (0) or ‘NOT NULL’ as true

iii. **FP Growth**

This operator is used to calculate all the frequent item sets occurring within a given data set. The FP means Frequent Patterns. Frequent item sets are simply groups of items which often appear together in data.

3.4.5. Evaluation Phase

In this phase, the results obtained were evaluated for the values and information produced. This section is discussed in detail in Chapter 4 where results and analysis of the study are outlined.

3.4.6. Deployment Phase

Once the models are complete, the solution developed will be handed over to the department concerned for re-use and extraction of new information from the database. The intention of giving a solution is the driving force of the whole process and a key deliverable of the deployment phase. It is hoped that the model will be most suited for use in similar scenarios.

Chapter 4

RESULTS AND ANALYSIS

4.1. Introduction

This chapter contains the results of the activities performed during the study that have been outlined in the Methodology phase. The chapter also gives an output of the results obtained with respect to the objectives the study had set out to achieve within the guidelines of the CRISP-DM framework.

4.2. Business and Data Understanding

The data source for the study is the output of the business process and workflow for the import and export permit issue. Figure 24 represents the workflow of the process of acquiring an import or export permit from PPB.

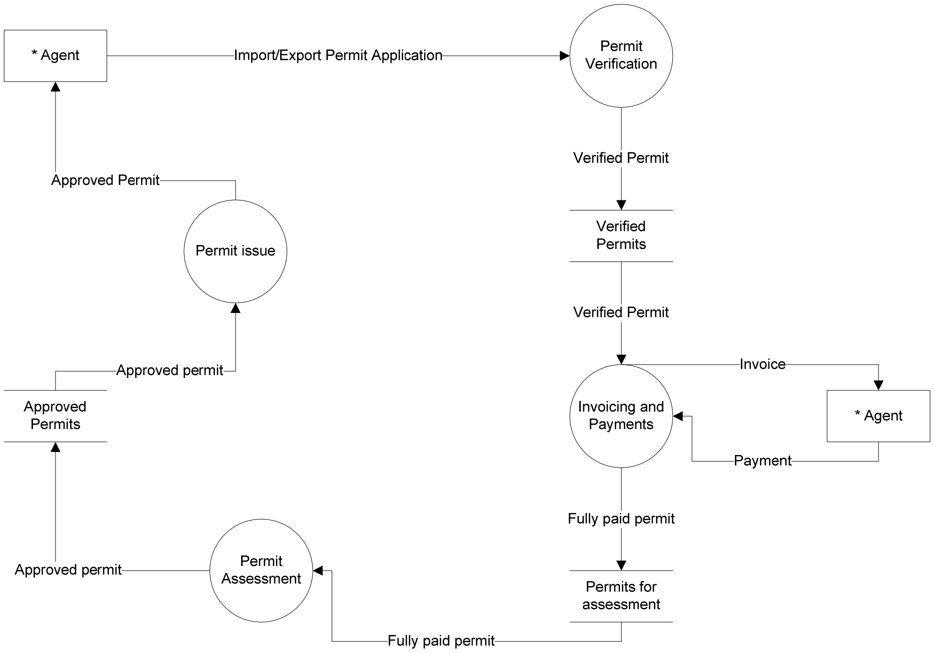
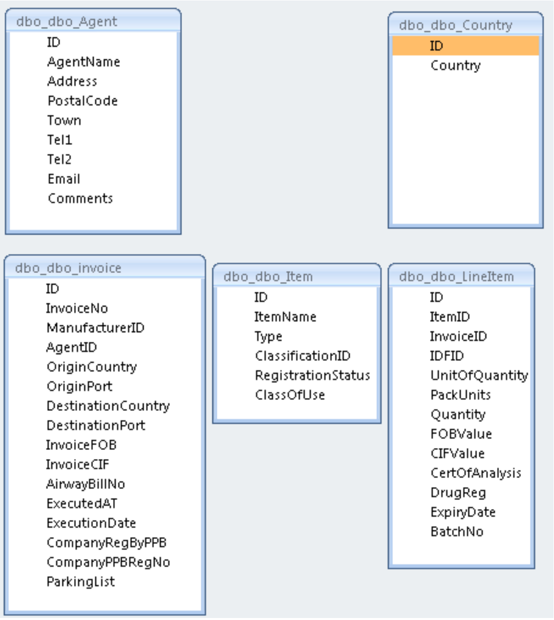


Figure 24: Data flow of the process of importation of a pharmaceutical product

The process is extensive although the area of concern to the study is the data store called "Approved Permits" from where the data for analysis is obtained. It is important to note that the data flow represents a manual system that has been in place to date and the data store "Approved Permits" is an archive of copies of the original permits issued to the agents who applied for them.

Currently, the electronic record of data contained in the permits goes back to the year 2008. The data is stored in Microsoft Access. Figure 25 shows a summary of how the Ms Access database was represented;



dbo_dbo_Agent
ID
AgentName
Address
PostalCode
Town
Tel1
Tel2
Email
Comments

dbo_dbo_Country
ID
Country

dbo_dbo_invoice
ID
InvoiceNo
ManufacturerID
AgentID
OriginCountry
OriginPort
DestinationCountry
DestinationPort
InvoiceFOB
InvoiceCIF
AirwayBillNo
ExecutedAT
ExecutionDate
CompanyRegByPPB
CompanyPPBRegNo
ParkingList

dbo_dbo_Item
ID
ItemName
Type
ClassificationID
RegistrationStatus
ClassOfUse

dbo_dbo_LineItem
ID
ItemID
InvoiceID
IDFID
UnitOfQuantity
PackUnits
Quantity
FOBValue
CIFValue
CertOfAnalysis
DrugReg
ExpiryDate
BatchNo

Figure 25: Database structure of Import Export Permit repository

4.3. Pharmaceutical Product Classification

Based on the raw data obtained from the data store of imports and exports, the pharmaceutical products were classified into product groups. This classification of pharmaceutical products is based on the Kenya Essential Drug list for consistency of naming.

Table 2: Pharmaceutical Product classification according to the Kenya Essential Drug List

Pharmaceutical Product Classification according to the Kenya Essential List
ANAESTHETICS
ANALGESICS ANTIPYRETICS, NSAIDs
ANTI-ALLERGICS & DRUGS USED IN ANAPHYLAXIS
ANTI-BIOTICS
ANTI-DIABETICS
ANTIDOTES & SUBSTANCES USED IN POISONINGS
ANTIEPILEPTICS
ANTI-FUNGALS
ANTI-MALARIALS
ANTIMIGRANE DRUGS
ANTINEOPLASTICS & IMMUNOSUPPRESSIVE DRUGS
ANTIPARKINSONISM DRUGS
ANTI-RETROVIRALS
ANTI-TBS
ANTI-VIRALS
BLOOD PRODUCTS & BLOOD SUBSTITUTES
CARDIOVASCULAR DRUGS
DERMATOLOGICAL DRUGS
DIAGNOSTICS AGENTS (RADIOLOGICALS)
DISINFECTANTS & ANTISEPTICS
DIURETICS
DRUGS AFFECTING BLOOD
GASTROINTESTINAL DRUGS
HORMONES, ENDOCRINE DRUGS, CONTRACEPTIVES
IMMUNOLOGICALS (VACCINES)
MISCELLANEOUS
MUSCLE RELAXANTS
OPHTALMOLOGICAL E.N.T. PREPARATIONS
OXYTOCICS & ANTIOXYTOCICS
PERITONEAL DIALYSIS SOLUTIONS
PSYCHOTHERAPEUTIC DRUGS
RESPIRATORY TRACT DRUGS
SOLUTIONS FOR WATER, ELECTROLYTE, ACID-BASE DISTURBANCE
VETERINARY PRODUCTS
VITAMIN & MINERALS

Analysis on the total value of products within the categories in Table 2 was done to obtain a graph that showed the most highly imported categories.

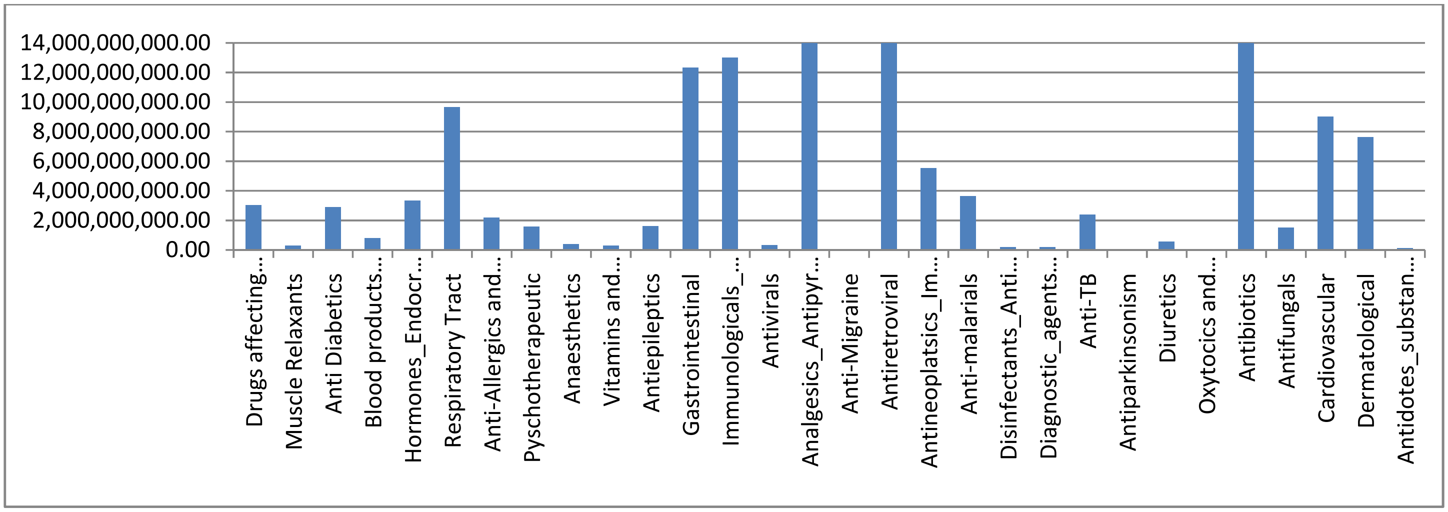


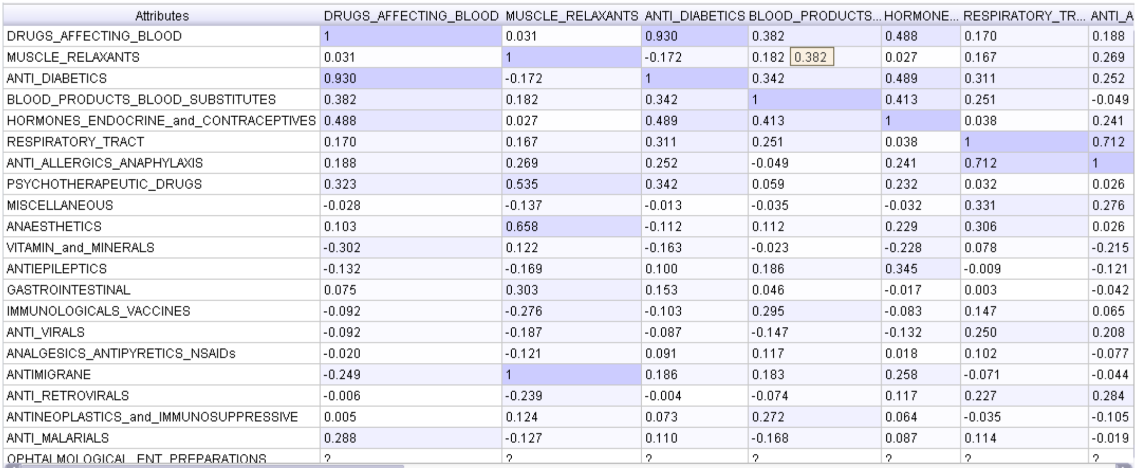
Figure 26: Summary of FOBValue of the various Categories of Pharmaceutical Products

The above product groups were then used to generate graphs of yearly imports of pharmaceutical products with “FOBValue” to observe the trend of each of the groups over time.

4.4. Correlation among Pharmaceutical Product Groups

Through correlation analysis, it can be seen that there are product groups which are related based on their FOB values. This is obtained from correlation analysis by Rapidminer that produces a Matrix indicating the level of correlation as shading of the coefficients obtained. The value ranges from 1 to -1 where a value approaching 1 implies the two groups are positively correlated while a value approaching -1 implies the two groups are negatively correlated. A sample of the matrix is as shown in Table 3. As seen from the matrix in Table 3, the pair wise correlations are very numerous and thus it is not possible to plot graphs for all the pairs of pharmaceutical product groups. For this reason, correlations were limited to product categories that passed the two billion mark as per Figure 4-3. Combined trend graphs of correlated pharmaceutical product groups were plotted to observe their correlations visually.

Table 3: Correlation Matrix showing the relationship between the various product groups



Attributes	DRUGS_AFFECTING_BLOOD	MUSCLE_RELAXANTS	ANTI_DIABETICS	BLOOD_PRODUCTS...	HORMONE...	RESPIRATORY_TR...	ANTI_A
DRUGS_AFFECTING_BLOOD	1	0.031	0.930	0.382	0.488	0.170	0.188
MUSCLE_RELAXANTS	0.031	1	-0.172	0.182 0.382	0.027	0.167	0.269
ANTI_DIABETICS	0.930	-0.172	1	0.342	0.489	0.311	0.252
BLOOD_PRODUCTS_BLOOD_SUBSTITUTES	0.382	0.182	0.342	1	0.413	0.251	-0.049
HORMONES_ENDOCRINE_and_CONTRACEPTIVES	0.488	0.027	0.489	0.413	1	0.038	0.241
RESPIRATORY_TRACT	0.170	0.167	0.311	0.251	0.038	1	0.712
ANTI_ALLERGICS_ANAPHYLAXIS	0.188	0.269	0.252	-0.049	0.241	0.712	1
PSYCHOTHERAPEUTIC_DRUGS	0.323	0.535	0.342	0.059	0.232	0.032	0.026
MISCELLANEOUS	-0.028	-0.137	-0.013	-0.035	-0.032	0.331	0.276
ANAESTHETICS	0.103	0.658	-0.112	0.112	0.229	0.306	0.026
VITAMIN_and_MINERALS	-0.302	0.122	-0.163	-0.023	-0.228	0.078	-0.215
ANTIEPILEPTICS	-0.132	-0.169	0.100	0.186	0.345	-0.009	-0.121
GASTROINTESTINAL	0.075	0.303	0.153	0.046	-0.017	0.003	-0.042
IMMUNOLOGICALS_VACCINES	-0.092	-0.276	-0.103	0.295	-0.083	0.147	0.065
ANTI_VIRALS	-0.092	-0.187	-0.087	-0.147	-0.132	0.250	0.208
ANALGESICS_ANTIPYRETICS_NSAIDs	-0.020	-0.121	0.091	0.117	0.018	0.102	-0.077
ANTIMIGRANE	-0.249	1	0.186	0.183	0.258	-0.071	-0.044
ANTI_RETROVIRALS	-0.006	-0.239	-0.004	-0.074	0.117	0.227	0.284
ANTINEOPLASTICS_and_IMMUNOSUPPRESSIVE	0.005	0.124	0.073	0.272	0.064	-0.035	-0.105
ANTI_MALARIALS	0.288	-0.127	0.110	-0.168	0.087	0.114	-0.019
OPHTALMOLOGICAL_ENT_PREPARATIONS	?	?	?	?	?	?	?

4.4.1. Visualization of Correlations

From the matrix in Table 3, it is possible to deduce that there is correlation of varying degree within the various categories of pharmaceutical products. Those with a darker background represent highly correlated pairs while the lighter the shade, the lower the degree of correlation.

The following are the selected correlations;

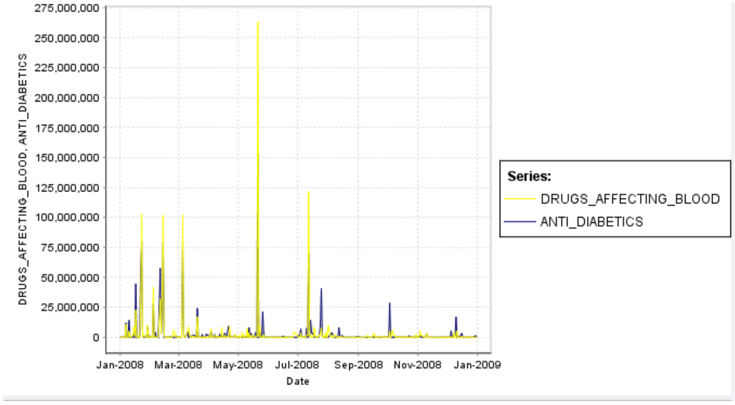


Figure 27: Trends of Drugs affecting Blood against Anti diabetics

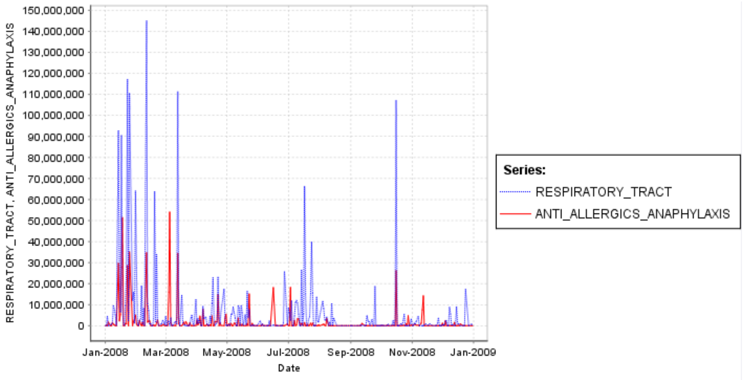


Figure 28: Respiratory Tract against Anti-Allergics

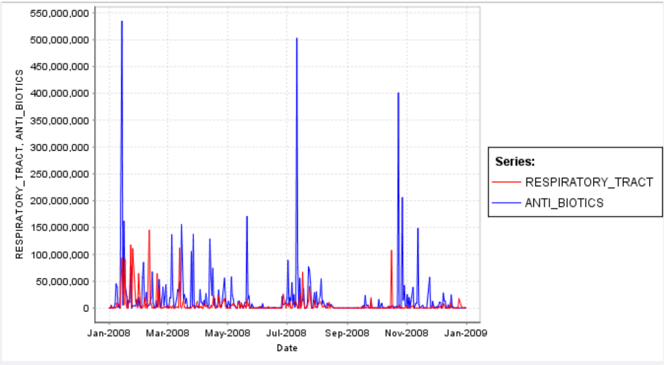


Figure 29: A graph of Respiratory Tract and Antibiotic drug FOB Values against time

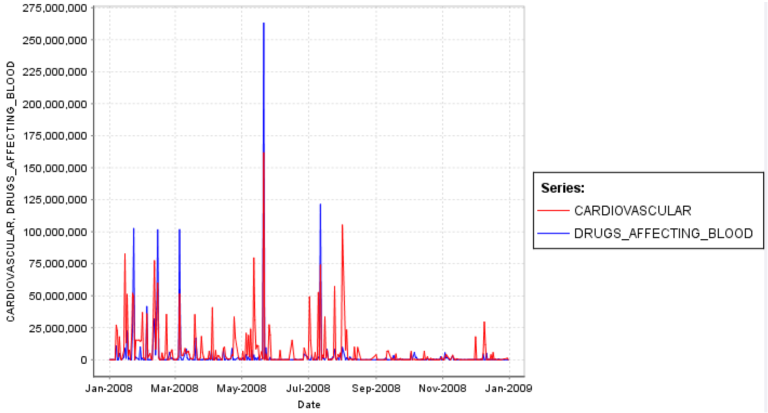


Figure 30: Drugs Affecting Blood and Cardiovascular Drugs

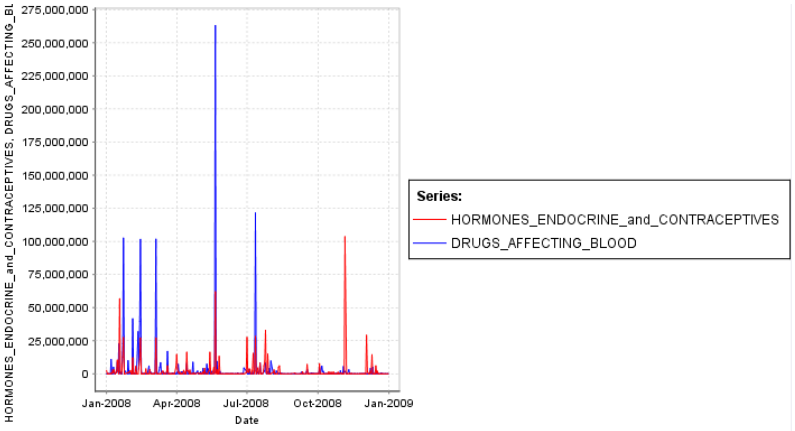


Figure 31: Drugs Affecting Blood and Hormones, Endocrine and Contraceptive drugs

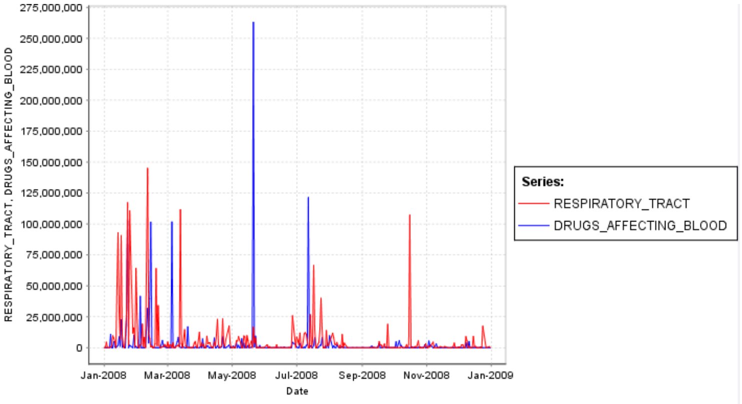


Figure 32: Drugs affecting Blood and Respiratory Tract Drugs

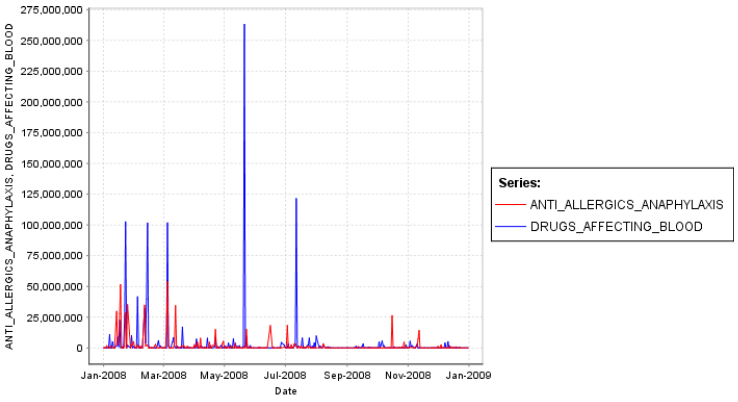


Figure 33: Drugs Affecting Blood and Anti-Allergics and Anaphylaxis

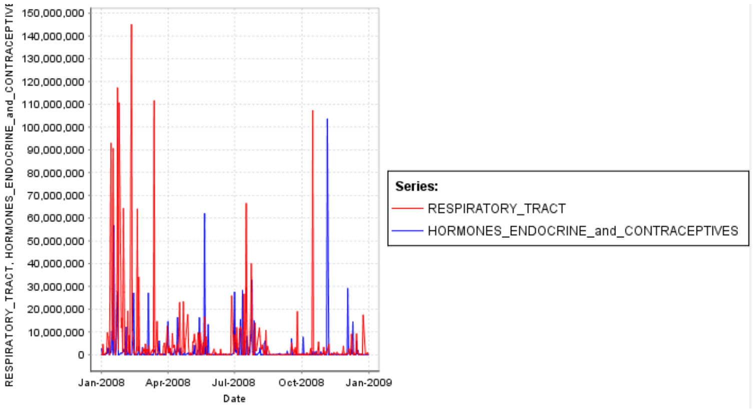


Figure 34: Respiratory Tract and Hormone, Endocrine and Contraceptives

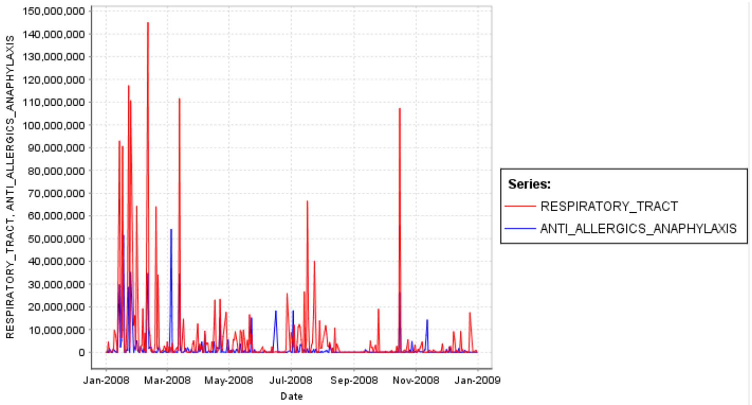


Figure 35: Respiratory tract and Anti-Allergics, Anaphylaxis

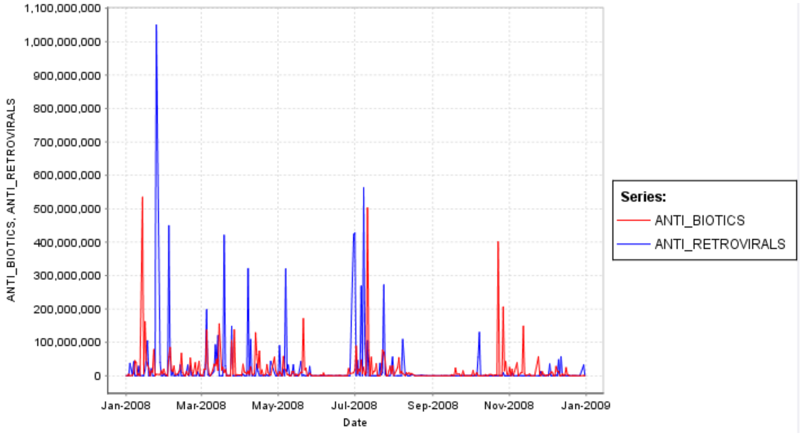


Figure 36: Antiretrovirals and Antibiotics

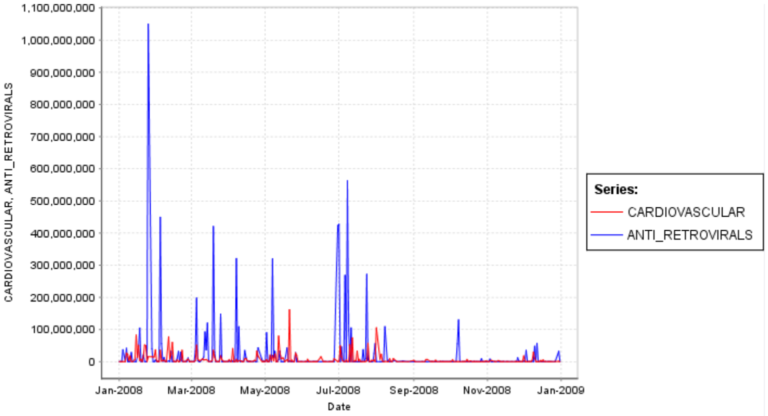


Figure 37: Antiretrovirals and Cardiovascular Drugs

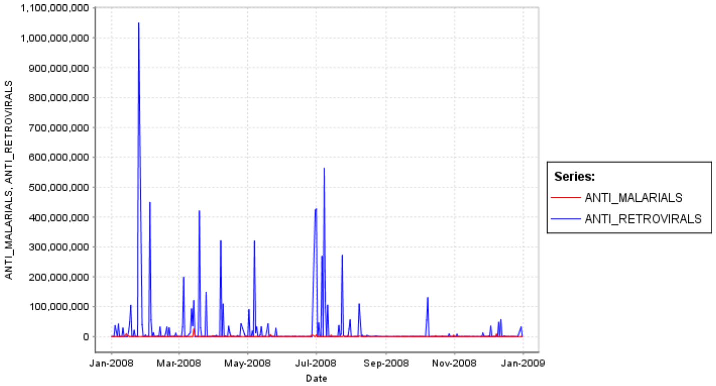


Figure 38: Antiretrovirals and Antimalarial Drugs

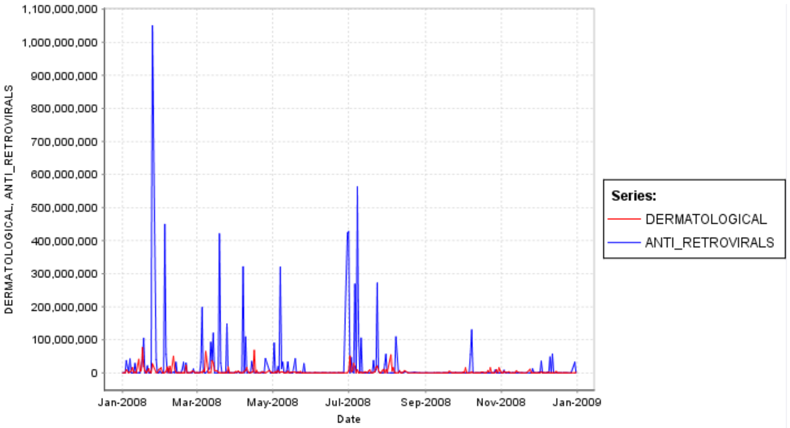


Figure 39: Antiretrovirals and Dermatological Drugs

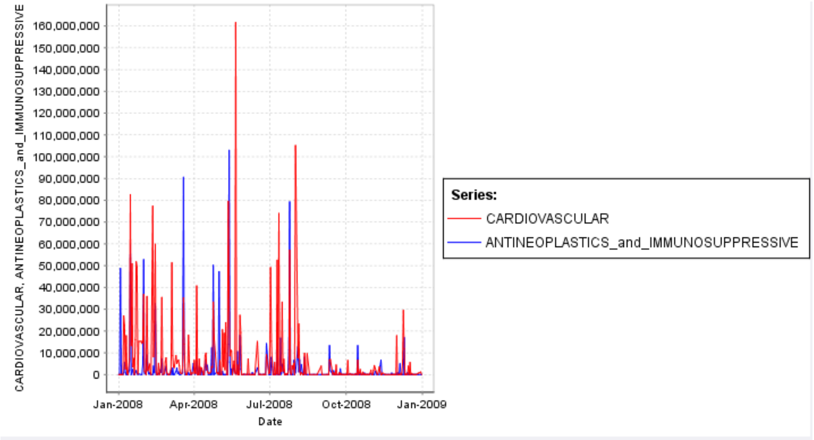


Figure 40: Cardiovascular and Antineoplastics, Immunorepressive

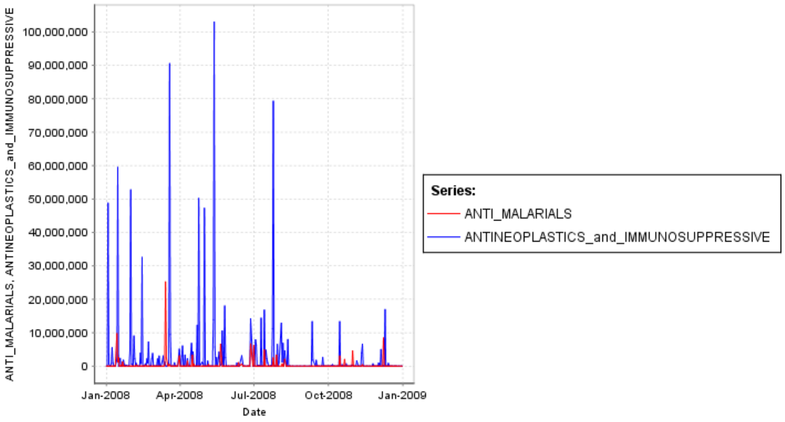


Figure 41: Antineoplastics_Immunorepressive and Antimalarials

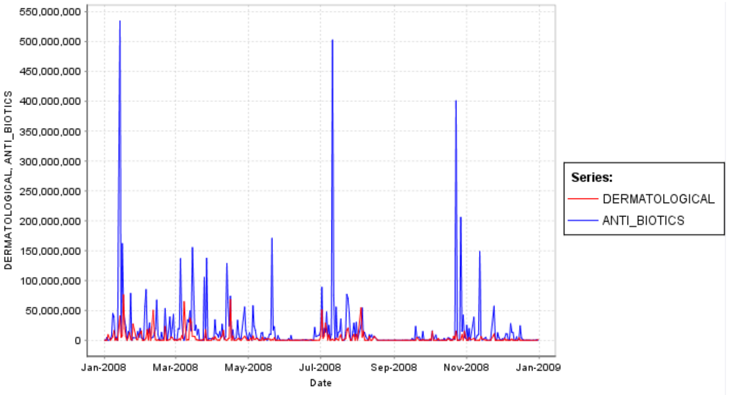


Figure 42: Antibiotics and Dermatological Drugs

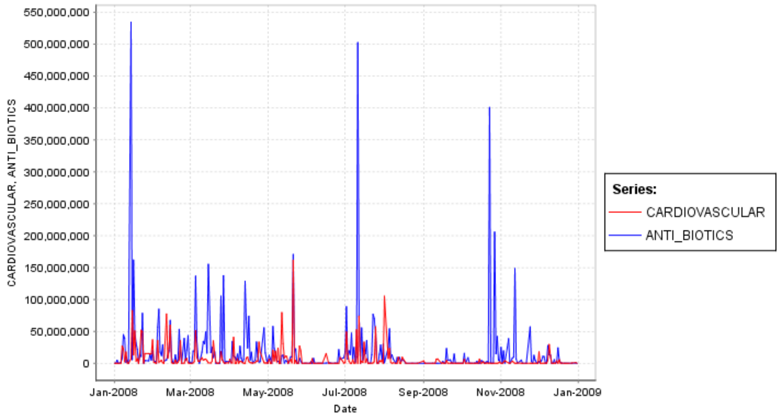


Figure 43: Antibiotics and Cardiovascular Drugs

From the selected plots of correlated product groups, the presence of correlation can be visually confirmed in addition to the correlation matrix.

4.5. Association Analysis

The occurrence of frequent item sets is common within transactions containing items. Frequent item sets are simply groups of items that often occur together in a data set. On analysis, the data yielded interesting associations. Table 4 shows the top ten associations by support obtained from analysis. It can be observed that the product group Antibiotics is associated to a number of other combinations of product groups which implies that within the various item sets, there exists an Antibiotic.

Table 4: Association Rules obtained from Rapid Miner at 0.95 confidence level

Premises	Conclusion	Support	Confidence
DERMATOLOGICAL	ANTI_BIOTICS	0.618557	0.952381
GASTROINTESTINAL, ANALGESICS_ANTIPYRETICS_NSAIDs	ANTI_BIOTICS	0.573883	0.95977
GASTROINTESTINAL, RESPIRATORY_TRACT	ANTI_BIOTICS	0.573883	0.965318
GASTROINTESTINAL, DERMATOLOGICAL	ANTI_BIOTICS	0.563574	0.982036
ANTI_BIOTICS, CARDIOVASCULAR	GASTROINTESTINAL	0.560137	0.958824
GASTROINTESTINAL, CARDIOVASCULAR	ANTI_BIOTICS	0.560137	0.970238
ANALGESICS_ANTIPYRETICS_NSAIDs, RESPIRATORY_TRACT	ANTI_BIOTICS	0.539519	0.969136
ANALGESICS_ANTIPYRETICS_NSAIDs, RESPIRATORY_TRACT	GASTROINTESTINAL	0.532646	0.95679
ANALGESICS_ANTIPYRETICS_NSAIDs, DERMATOLOGICAL	ANTI_BIOTICS	0.52921	0.968553
ANALGESICS_ANTIPYRETICS_NSAIDs, CARDIOVASCULAR	ANTI_BIOTICS	0.522337	0.962025

Plots to evaluate the above associations are shown in figures 44 to 52;

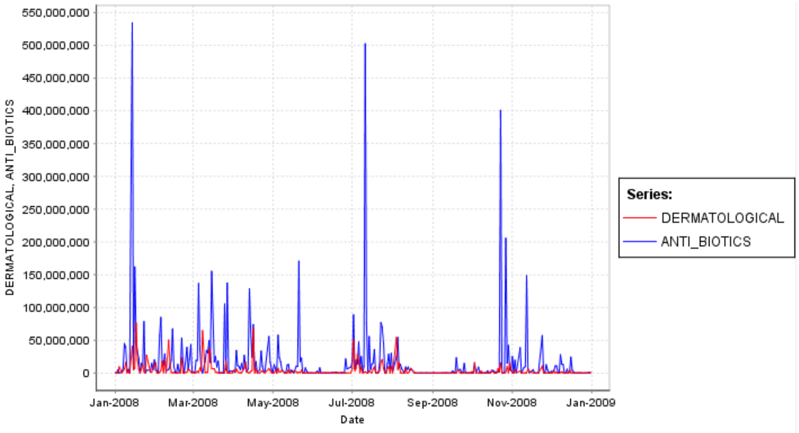


Figure 44: Dermatologicals and Antibiotics

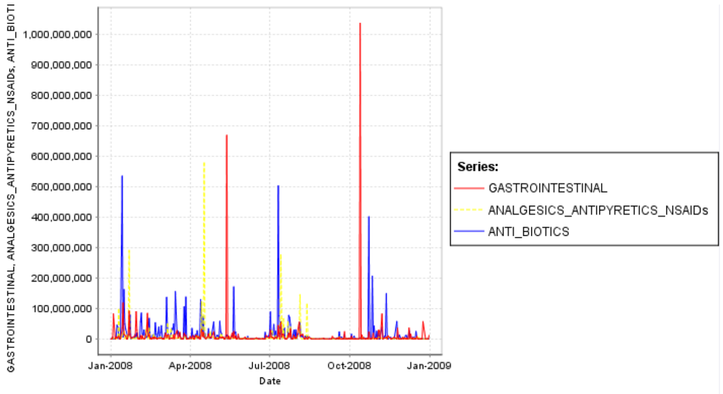


Figure 45: Gastrointestinal, Analgesics_Antipyretics_NSAIDS and Antibiotics

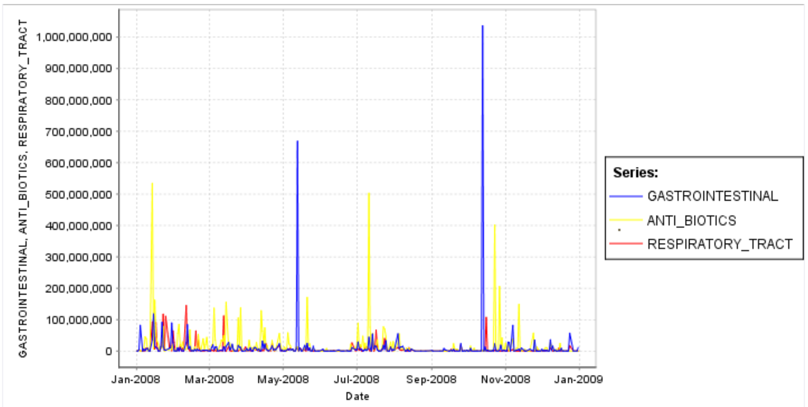


Figure 46: A graph evaluating the association among three product groups

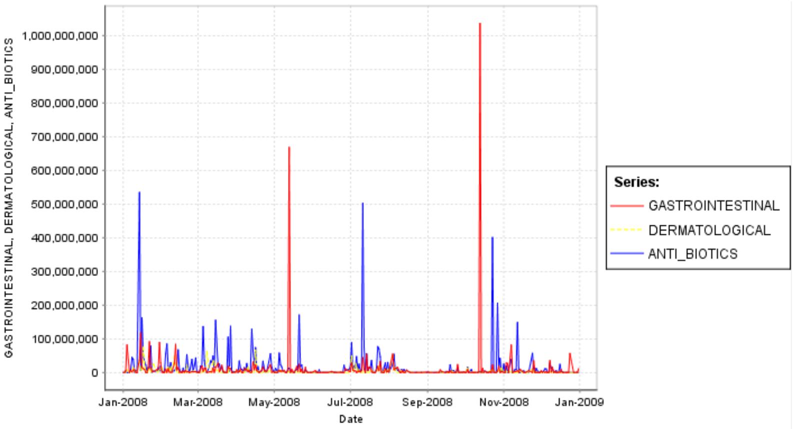


Figure 47: Gastrointestinal, Dermatologicals and Antibiotics

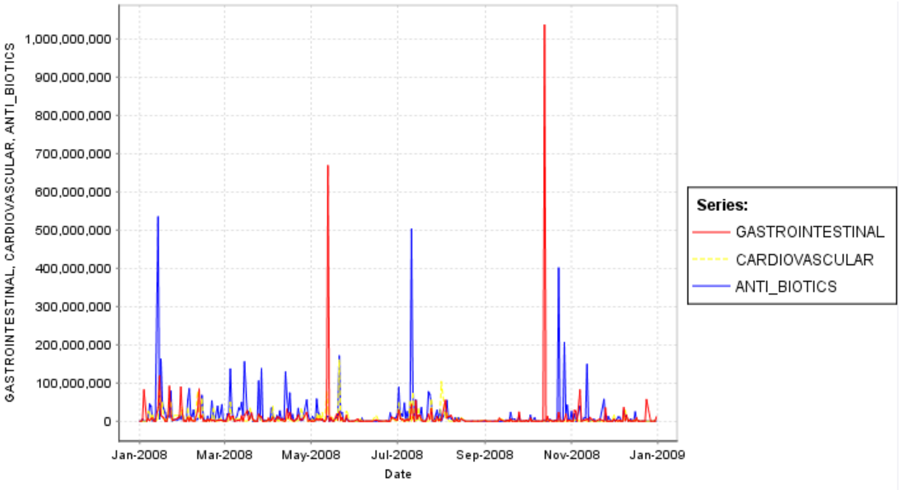


Figure 48: Antibiotics, Cardiovascular and Intestinal Drugs

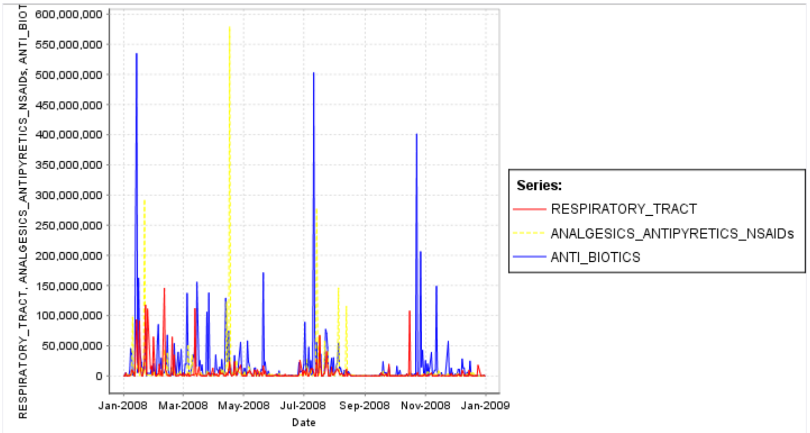


Figure 49: Analgesics_Antipyretics_NSAIDS, Respiratory Tract and Antibiotics

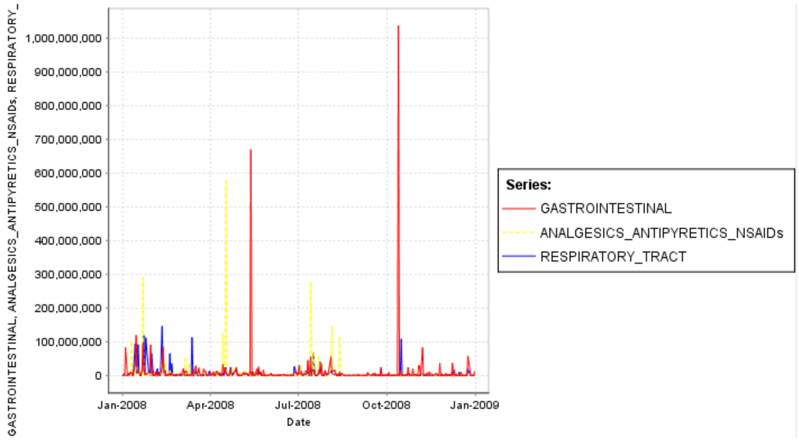


Figure 50: Analgesics_Antipyretics_NSAIDS, Respiratory Tract and Gastrointestinal Drugs

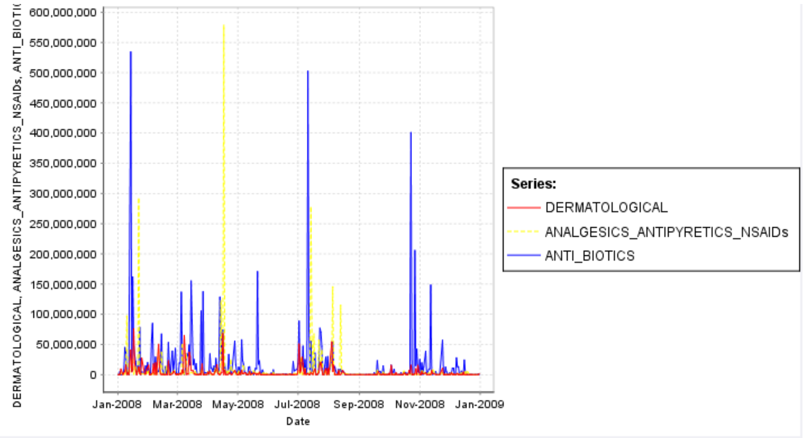


Figure 51: Analgesics_Antipyretics_NSAIDS, Dermatological and Antibiotics

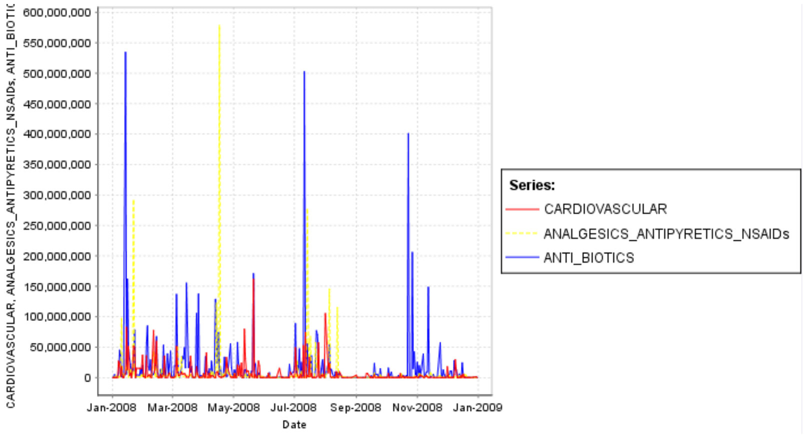


Figure 52: Analgesics_Antipyretics_NSAIDS, Cardiovascular and Antibiotics

4.6. Conclusion

After performing correlation and association analysis on the data on imports of pharmaceutical products, some inferences can be made from it with regards to the information illustrated from the graphs obtained. From correlation analysis, it can be concluded that within the pharmaceutical product groups imported, there are related pairs based on the results of the coefficients in the correlation matrix.

From association analysis, it can be shown that there are associations between several product groups which are represented as frequent item sets. These are indicative of various product groups being imported at the same time. Graphs of combined product groups against time for example GASTROINTESTINAL, RESPIRATORY_TRACT, ANTI_BIOTICS shows a similar trend to confirm their association with periods of inactivity in June and September. The following is a brief description of the functions of the mentioned categories;

Gastrointestinal

These are drugs that are used to treat ailments of the digestive system which is also referred to as the gastrointestinal tract

Antibiotics

These are drugs that are used to treat diseases brought about by bacterial infections.

Respiratory Tract

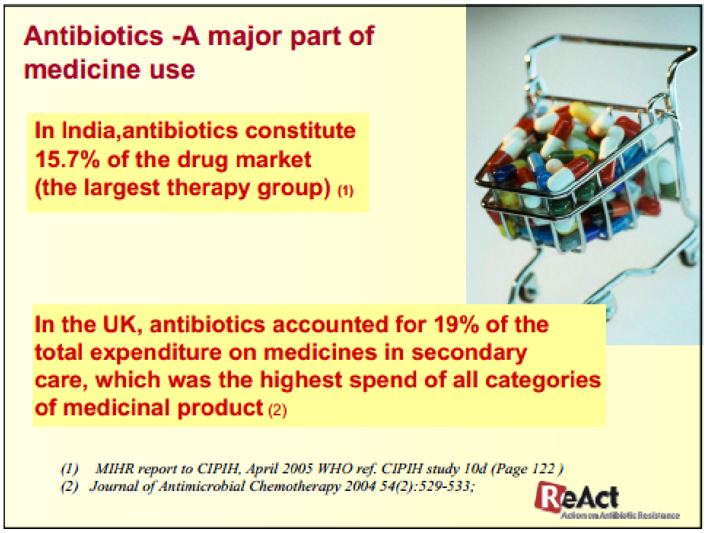
These are drugs that are used in the treatment of diseases affecting the respiratory system.

There is a major concern on the use of antibiotics, cough and cold medicines, painkillers and anti-diarrhoeals in many developing countries (Le Grand, Hogerzeil, & Haaijer-Ruskam, 1999). The study also shows that the sales of the aforementioned classes of medicines exceed the medical condition that they are supposed to treat.

Table 5: Summary of the number of drugs contained per prescription

Prescription Containing	Number of prescriptions	Percentage
One	2	0.67%
Two	10	3.33%
Three	89	29.67%
Four	158	52.70%
Five	34	11.30%
Six	6	2%
Seven	1	0.30%
Total	300	100%

Source: Audit of Prescribing Practices to Evaluate Rational Use of Medicines in the OPD of Orthopaedics in a Private Medical College Hospital.



Antibiotics -A major part of medicine use

In India,antibiotics constitute 15.7% of the drug market (the largest therapy group) (1)

In the UK, antibiotics accounted for 19% of the total expenditure on medicines in secondary care, which was the highest spend of all categories of medicinal product (2)

(1) MIHR report to CIPIH, April 2005 WHO ref. CIPIH study 10d (Page 122)
(2) Journal of Antimicrobial Chemotherapy 2004 54(2):529-533;

Figure 53: General Trend of Medicine use

Chapter 5

CONCLUSION

5.1. Introduction

This chapter contains the summary of the study. It contains the conclusions made from the study in terms of knowledge obtained and added to the common field of knowledge. This chapter also looks into the hurdles and limitations encountered during the study and propose recommendations towards future work.

5.2. Findings of the Research

From this research, it can be observed that correlations and associations are present among the various pharmaceutical product groups. There correlations and associations are well illustrated through the observation of trends of the particular product groups of interest. These trends likely occur due to the market responding to a particular need or situation.

The situation in developing countries in terms of pharmaceutical product use is generally the same thus there are similar trends of pharmaceutical product use in countries such as Bangladesh, Ethiopia, Nigeria, Uganda and Tanzania. Research has shown that a visit to a hospital is likely to result in the prescription of a number of medicines to treat or control the ailment or condition suffered. Usually quite a number of Antibiotics are prescribed (Adebayo & Hussain, 2010).

Hospital Research in Bangladesh on drug use and prescribing patterns of medical staff shows that the trend is on issuance of multiple medicines or poly-pharmacy (Afsan et al, 2012). The same phenomenon is replicated in Ethiopia whereby the average number of drugs per prescription ranges from 1.98 to 2.24 (Angamo, Wabe, & Raju, 2011). The issuance of multiple medicines to a patient implies that there are associations between the various types of medicines available in the country and hence the importance of discovering these association patterns within them.

5.3. Contribution to the Pharmacy and Poisons Board

It is hoped that this research will go a long way in giving the regulator a better insight into the permits they issue to agents for imports and exports of pharmaceutical products. With regards to the association between the various pharmaceutical product groups and the deduction that it is linked to actual use, then PPB might enforce its influence in the use of doctors' prescriptions in drug shops or liaise with partner regulatory boards such as those regulating doctors to look into the composition of their prescriptions of medicines to patients. The study may provide PPB with indications of excessive or limited amounts of a particular category of product so as to trigger the necessary regulatory action within its mandate to control the situation and avert negative effects to the country’s citizens. The study might also trigger the investigation of the use of a particular product group for example Antibiotics with the aim of enforcing their proper use to reduce resistance and other issues related to their abuse.

5.4. Limitations of the study

The study was limited by the pace at which the data entry team at the PPB which is currently heavily engaged in the process of making electronic its hard copy documents regarding imports and exports. Another issue was the records that were missing several important variables such as product names, dates and cost which would have been valuable to the study.

5.5. Recommendations for future work

Knowledge obtained from data mining of medical prescription data paired with that of imports and exports would be quite interesting as it would bring to light the accurate consumption of data of pharmaceutical products in the country as this is still a grey area in Kenya. PPB is in the process of implementation and enforcement of consumption reporting system to all the distributors and wholesale outlets in the country and data mining of the newly generated data would provide great insight into movement of pharmaceutical products in the country. The fact that data will now be generated by an automated system rather than data entry of historical records, this will provide relatively clean data for analysis of pharmaceutical products. With regards to the association between the various pharmaceutical product groups and the deduction that it is linked to actual use, then PPB might enforce its influence in the use of doctors’ prescriptions in drug shops or liaise with partner regulatory boards such as those regulating doctors to look into the composition of their prescriptions of medicines to patients. The study may provide PPB with indications of excessive or limited amounts of a particular category of product so as to trigger the necessary regulatory action within its mandate to control the situation and avert negative effects to the country’s citizens. The study might also trigger the investigation of the use of a particular product group for example Antibiotics with the aim of enforcing their proper use to reduce resistance and other issues related to their abuse.

References

1. Adebayo, E. T., & Hussain, N. A. (2010). Pattern of prescription drug use in Nigerian army. *Annals of African Medicine* , 152-158.

2. Afsan, M., Haque, M., Alam, M., & Noor, N. (2012). Audit of Prescribing Practices to Evaluate Rational Use of Medicines in. *J Shaheed Suhrawardy Med Coll* , 39-42.

3. Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M. J., Ventura, S., Garrell, J., et al. (2011). KEEL: a software tool to assess evolutionary algorithms for Data Mining Problems. *Soft Computing* , 307–318.

4. Angamo, M. T., Wabe, N. T., & Raju, N. J. (2011). Assessment of Patterns of Drug use by using World Health Organization's Prescribing,Patient Care and Health facility indicators inSelected Health Facilities in Southwest Ethiopia. *Journal of Applied Pharmaceutical Science* , 62 - 66.

5. Berzal, F., Cubero, J.-C., Marín, N., Sánchez, D., Serrano, J.-M., & Vila, A. (2005). Association rule evaluation for classification purposes. *Actas del III Taller Nacional de Mineria de Datos y Aprendizaje,* , 135-144.

6. Džeroski, S. (2007). Towards a General Framework for Data Mining. *Knowledge Discovery in Inductive Databases* .

7. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2012). The WEKA Data Mining Software: An Update. *SIGKDD Explorations , Volume 11* (Issue 1), 11 - 18.

8. IBM Corporation. (2012). *IBM SPSS Modeler*. Retrieved from IBM Software: http://public.dhe.ibm.com/common/ssi/ecm/en/ytd03124usen/YTD03124USEN.PDF

9. Kajungu, D. K., Selemani, M., Masanja, I., Baraka, A., Njozi, M., Khatib, R., et al. (2012). Using classification tree modelling to investigate drug prescription practices at health facilities in rural Tanzania. *Malaria Journal* , 11: 311.

10. Kantardzic, M. (2011). *Data Mining: Concepts, Models, Methods, and Algorithms.* New Jersey: John Wiley & Sons.

11. KNIME.com AG. (n.d.). Retrieved from http://www.knime.org/files/Marketing/Datasheets/KNIME_Desktop_PDS.pdf

12. KNIMEtech. (n.d.). *Features*. Retrieved from KNIME : http://www.knime.org/features

13. Kriegel, H.-P., Borgwardt, K. M., Kröger, P., Pryakhin, A., Matthias, & Zimek, A. (2007). Future trends in data mining. *Data Mining Knowledge Discovery* , 15:87–97.

14. Kumar, P., Sehgal, V. K., Sehgal, N. K., & Chauhan, D. S. (2012). A Benchmark to Select Data Mining Based Classification Algorithms for Business Intelligence and Decision Support Systems. *International Journal of Data Mining & Knowledge Management Process* .

15. Kurgan, L. A., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining Process Models. *The Knowledge Engineering Review* , 1 - 24.

16. Kusiak, A., & Shah, S. C. (2006). Data Mining and Warehousing in Pharma Industry. *Encyclopedia of Data Warehousing and Mining, Idea Group, Inc* , 239 - 241.

17. Le Grand, A., Hogerzeil, H. V., & Haaijer-Ruskam, F. M. (1999). Intervention Research in Rational use of Drugs: a Review. *Health Policy and Planning* , 89 - 102.

18. Mikut, R., & Reischl, M. (2011). Data mining tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* , 1 - 13.

19. Odongo, M. (2012). Adoption of Open Source and Open Standards in Academic Libraries in Kenya. *Proceedings from the 20th Standing Conference of Eastern, Central and Southern Africa Library and Information Associations.* Nairobi.

20. Okoro, R. N., & Shekari, B. G. (2013). Physicians' drug prescribing patterns at the national health insurance scheme unit of a teaching hospital in the North Eastern Nigeria. *Archives of Pharmacy Practice* , *4* (1), 3-8.

21. Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng* , 23(4), 3-13.

22. Ranjan, J. (2007). Application of Data Mining Techniques in the Pharmaceutical Industry. *Journal of Theoretical and Applied Information Technology* , 61 - 67.

23. Saxena, K., & Rajpoot, D. (2009). A Way to Understand Various Patterns of Data Mining Techniques for Selected Domains. *International Journal of Computer Science and Information Security* , *Volume 6*, 186 - 191.

24. Sharma, S., Osei-Bryson, K.-M., & Redmond, R. (2008). *An Integrated Knowledge Discovery and Data Mining Process Model.* Virginia.

25. Silwattananusarn, T., & Tuamsuk, K. (2012). Data Management and Its Application for Knowledge Management: A Literature Review from 2007 to 2012. *International Journal of Data Mining & Knowledge Management Process* , 13 - 24.

26. Sowan, B. I. (2011). *Enhancing Fuzzy Associative Rule Mining Approaches for Improving Prediction Accuracy.* PhD Thesis, University of Bradford, School of Computing, Informatics & Media, Bradford.

27. United Nations. (2013). *Narcotic Drugs Estimated World Requirements for 2013 (Statistics for 2011).* New York: United Nations Publication.

28. Viktil, K. K., Blix, H. S., Moger, T. A., & Reikvam, A. (2006). Polypharmacy as commonly defined is an indicator of limited value in the assessment of drug-related problems. *British Journal of Clinical Pharmacology* , 187–195.

29. Wahbeh, A. H., Al-Radaideh, Q. A., Al-Kabi, M. N., & Al-Shawakfa, E. M. (2011). A Comparison Study between Data Mining Tools over some Classification Methods. *International Journal of Advanced Computer Science and Applications* (Special Issue on Artificial Intelligence), 18 - 26.

30. Wang, Y., & K. C. Wong, A. (2003). Pattern Discovery: A Data Driven Approach to Decision Support. *IEEE Transactions on Systems, Man, and Cybernetics* .

31. Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* , 29 - 39.

32. Zhang, S., Zhang, C., & Yang, Q. (2003). Data Preparation for Data Mining. *Applied Artificial Intelligence* , 375 - 381.