

Table of Contents

Section One - Data-Set Structure and Scope	3
Introduction	3
Data Mining Concepts.....	3
Data-Set Description	4
Source of Data.....	4
Data Purpose.....	4
Structure	5
Pre-Processing.....	6
Extraction	6
Post-extraction Pre-Processing	7
Pattern Discovery.....	10
Classification	12
Clustering	13
Association.....	15
Secondary Pre-Processing.....	16
Attribute Removal.....	16
Analyses and Findings	17
Reflection	17
References	19
Appendix A.....	20
Extraction Script.....	20

Figures

Figure 1: Field Definitions	5
Figure 2: NULL Value Handling.....	7
Figure 3: Change of Date.....	8
Figure 4: Removal of Records	9
Figure 5: Grouping	9
Figure 6: Initial Load Visualisation	10
Figure 7: Post-Discretisation View	11
Figure 8: Scheduled_Ship_Date Classification Error	12
Figure 9: Location RandomTree Error.....	12
Figure 10: Category Random Tree Classification	13
Figure 11: Scheduled_Ship_Date KMeans	14
Figure 12: Location Clustering Representation.....	14
Figure 13: Category Clustering.....	15
Figure 14: Secondary Attribute Removal.....	16

Section One - Data-Set Structure and Scope

Introduction

In most operational companies, a very large amount of data passes through the organisation every day in various different forms, some of this data is used primarily for retrieval of customer or client information, some is updated to keep records current and other data input for information stored on new clients. With the amount of information that is exchanged, it would be logical to assume that a company would know everything there is to know for their target audience, though this is not always the case, there are many disparate efforts to ascertain such knowledge. One of the methods employed to understand the habits and purchasing processes as exhibited by their clientele is the process of data mining.

Data-mining is a term attributed to the retrieval of meaningful information from the masses of data that are available through both historic and current data, with more of an emphasis on patterns coined from historic data. This process allows for an organisation to view an abstracted representation of such historic data and as such provide an educated decision as to the best direction in which to steer the company. If for example, a manager was able to see that a particular products demand has increased at certain times during the year for the past several years, then he may decide to increase the available stock for this period to ensure that supply of the item met the demand for it. (1)

Data Mining Concepts

There are four major areas where the use of data mining is considered useful for business and can provide intelligence to managers to better inform decisions on current or future business practices, these include; classification, estimation, prediction and description. All of the aforementioned uses provide a wealth of possible business application and as such contribute to the construction of beneficial intelligence. Such uses also extend into several areas of businesses each suited to particular business types. (2)

The term classification in this context is used to denote the segregation and grouping of data records to construct a model by which all entities within the given data set may be identified. This concept allows for the division of the records based on one or many of the composite attributes. The use of classification is more popular within industries such as the insurance sector. An example of such classification being utilised would be the way in which a company may classify the client as a safe driver or potentially risky driver. It is based upon this classification that insurance prices may be determined. (3)

In addition to the use of classification a company may wish to utilise the collected data to extrapolate possible value results. This is often the case where no value is present therefore other means need to be followed in order to obtain such information, this is achieved through the process of estimation. Estimation allows for the calculation of a value based upon previous items with the same affecting factors and attributes. One such example of this would be in the form of building value.

Another use for data-mining within industry is for the prediction of certain recurrent trends and patterns. Prediction is used mostly in organisations that require reasonably accurate data upon which they may base an ideology for a future outcome. This approach is common in analysing buying habits and rises and falls in the purchase of products.

The fourth use of data-mining is far less computationally intensive and as such is far more easily obtained and used. The description provided by the use of this information is often used to providing profiling characteristics to the field in which they belong.

Data-Set Description

To illustrate the use of data-mining on vast data collections and provide an outline for the procedure outlined above, the following sections will provide a detailed explanation of the steps that are required to derive meaningful information from such historic data. Within this scenario, several of the key concepts employed within the field of data mining will be covered giving detailed explanation of their purpose and how to implement them. As a result of this process, it is believed that the benefits of such a practice and the data utilised within will be realised.

To perform the aforementioned task, the first requirement for the proposed scenario is the sourcing of a suitable collection of data upon which to base the investigation. To ensure that the scenario follows the theme of a realistic business process, several constraints were imposed upon the data collection to be utilised. To perform some of the tasks necessary, a dataset of a suitable size is important, attempting to perform analyses on smaller data sets typically below a one thousand records, the process of data-mining is almost rendered redundant. As a result of the previous note, the data set collected provides a range of roughly twenty eight thousand records, providing for a viable repository of data to perform said analyses upon.

Source of Data

The data viewed in the proposed scenario, was retrieved from an operational sales process within an organisation who sell products relating to office-ware including stationary and furniture. The company has been in existence for over a decade and although continue to expand, have no real means of targeting their products to specific clientele or the level of demand that they have for specific products. The sales data that has been amassed, includes elements relating to the order itself with all associated dates and times and also provides an insight into the items related to said order.

As this data is used within every day operations, it is not possible to perform this analysis on real-time data therefore a snapshot of the data has been created within a specific time period ranging between the years 2013 and 2014. This provides for a large enough dataset with the required attributes. If an extract larger than this were to be utilised there would be far too many records to accurately represent the results. (4)

Data Purpose

In the previous section, the creation source for the data was briefly covered. This background tendered an insight into the process that creates the information upon which the investigation is being conducted. It is the function of this section to provide a more detailed description of what this data is used for.

When a customer elects to purchase an items from a vendor, a specific well-defined sequence of events is followed, there may be several variations as per the policy of the vendor, however in many case the procedure is universal. A customer opts to purchase an item, a record is created to service this request alongside any products related to this order, the items are retrieved from the warehouse and placed within its packaging, a shipping record is created and the item is marked as dispatched. The customer may at this point perform tracking functionality or other checking that is permitted by the vendor or retailer. This procedure is the one followed in the creation of the data for the data set.

Structure

The data set acquired, was extracted from a database located on an SQLServer system and as such has a well-defined data storage structure, providing a somewhat cleaner collection of values than those that may be retrieved from other systems such as written documents and or images. The data was extracted through use of a custom written script to handle abnormalities within certain data elements, each of which are noted within the pre-processing section below along with the methods employed to mitigate the issues. The script utilised to extract the information may be viewed within appendix A.

The structure of the data has been outlined below, allowing for the clear visibility of all of the attributes concerned with the specified data type for each field. The importance of the fields has not been discussed within this section however will be covered in later sections of this text. It is important to note the data types of the fields in this way as it makes for a reference point for the pre-processing phase. The below structure denotes the data as stored within a comma separated value file (CSV) before any act of pre-processing other than through the retrieval process.

Attributes

The attributes are the elements of data within each of the records, of which there are 22 for the specified data set. There is a large range of data types included within this excerpt each with particularly suited purposes. The information provided represents the original data set before any pre-processing.

Figure 1: Field Definitions

Field Name	Data Type	Description
Order_No	Numeric	This number represents the unique identification number as created for the order record, this is then referenced by the order line record to attribute product details to the order.
Ext	Numeric	This number provides a means of identifying whether the record relates to an extension to an order, a value of 1 in this field indicates that the order is not an extension.
Cust_Code	Alphanumeric	This field stores a customer code value as stored within a reference table of the source database.
Ship_To	Alphanumeric	Provision of a customer shipping address code, orders may be shipped to addresses other than where the purchase was actioned from.
Req_Ship_date	DateTime	This field holds a date value indicating the latest date that the order can be shipped if it is to meet its requested delivery date.
Sch_Ship_Date	DateTime	The date that the specified order is to be shipped according to the shipping schedule.
Date_Shipped	DateTime	The date the actual shipment occurred may be represented by a NULL value if no shipment has occurred.
Date_Entered	DateTime	This is the date that the order was entered onto the data storage system.
Cust_PO	Alphanumeric	Provides posting code information (internal posting code not national standardised posting code).
Who_Entered	Textual	The name of the employee who entered the order in to the system.
Status	Character	A single character indicating the status of the order (V=Void,P=In transit,invoiced,Q=Quote,A=Awaiting Approval,S=Shipped).
Phone	Alphanumeric	Indicates the phone contact attributed to the account, this

		element does not provide consistent data due to a system update having caused a bug within data storage for this field.
Line_No	Numeric	This value indicates the index of the item line on the order.
Location	Textual	Provides indication of the location of the item within the organisation.
Category	Alphanumeric	Provides the internally recognized category code to which an item.
Time_entered	DateTime	The time that the line was entered for the order. Lines may have been entered at differing times to the overall order due to amendments.
Ordered	Numeric	The quantity of the product ordered for the specified line
Shipped	Numeric	The quantity already having been shipped for the specified line.
Price	Decimal	The price the customer pays for the item associated with the specified line.
Cost	Decimal	The cost of the item to the company for the specified line item.
Delivery_Date	Numeric	The delivery date of the items for the line in Julian format.
Sch_Ship_Date	Numeric	The scheduled ship date for the line item in Julian format.

The attributes noted provide a representation of the level of detail that is provided within the data set retrieved from the data store. It is possible to identify from this description of the attributes what data is held about the unique elements of the process. There are a number of differing approaches that may be followed to produce meaningful and useful information from this, however there are some pieces of data that simply have no value when attempting to mine. These elements have been identified and consequently removed within the pre-processing phases outlined in a proceeding section of this text.

A number of the fields outlined in figure1 contain NULL data or data with inconsistencies within the content, the possibilities for which have been noted within the appropriate fields, however this notation should not be taken as a complete reference to the fields that will contain such values, this may occur in other fields also. Where possible the permissible finite values have been indicated for those fields that contain them and a small summary of the meaning of said values has been provided.

Pre-Processing

The majority of data constructed or retrieved for data-mining from a business process almost always contains at least some element of noise, the proposed data store is no different from this and as such requires sanitisation. Sanitisation of the data is achieved through pre-processing. Pre-processing is the act of manipulating the representation of data, the removal of records and grouping of data to make the mining of said data far easier. (5)

Extraction

The origin of the data is such that manipulating the representation of the data is relatively easy in comparison to other formats. The use of SQL to format the data output from the server made the initial stages of data sanitation far easier than if the data had to be extracted manually. As some of the records contain attributes with NULL value data, it was believed important to deal with this task first.

NULL values present within the data cannot be utilised for effectively extrapolating useful information from the raw data, it is therefore necessary to either remove or substitute such data for a known value. The removal of the records that contained such entities would reduce the record set

however would also reduce the certainty with which the resultant information could be stated. The substitution of a NULL value in the data for a numeric value of 0 would retain the present level of records and provide a known numeric value with which to demonstrate bias in the dataset.

In contrast to this however, some of the attributes are of a data type other than standard numeric, some including date time values and alphanumeric characters. The presence of such anomalies, pose for an issue with the previous approach in that the tool used for the data mining process does not cope well with the introduction of empty strings. With this in mind it is evident that it is not possible to substitute the NULL values within alphanumeric fields to be the value of an empty string. It is possible however to substitute such values with a known value including “N/A” symbolising the data is not present or not applicable.

Where the field is of type DateTime and contains a NULL value a substitution has been made for the maximum possible date as provided by SQL Server at time of constructing this document (31 Dec 9999 23:59:59:000). This serves to provide a known value also indicating that such values are relative in that these records are ones that have not been correctly entered providing yet more avenues of possible intrigue for mining at a later stage. The processes outlined regarding substitution are specifically limited to NULL values but will be utilised where necessary for empty values or those with junk data, predominantly the use of empty strings within fields

The majority of the extraction process was achieved through implementation of an SQL script which was then executed to perform the extraction. The full script may be viewed within Appendix A, however some of the more relevant elements have been outlined here especially those relating to the aforementioned substitution of data.

Figure 2: NULL Value Handling

```
ISNULL(sch_ship_date, '31 dec 9999 23:59:59:000')  
ISNULL(ordered, 0)  
ISNULL(status, v)
```

Figure 2 illustrates the use of the ISNULL() function provided by SQL server and thus handling values that are empty or nonsensical. It is believed necessary to mark all those order without an appropriate status as void as their original status is unknown.

Post-extraction Pre-Processing

Following the extraction process, and the formatting that was enacted as part of it, it is important to perform some element of pre-processing to ensure data uniformity. To better explain this, view the example in figure 3. The first item displays a date time value with only the date value and the second item displays an item from the same field with a time value. This is treated as an inconsistency and causes many problems when attempting to provide a date format for the field within the ARFF file to be input into WEKA. (6)The following sections outline the functions performed within this process and provide explanation for each.

Data Transformation

As disparities exist between values within the same field it is consequently necessary to provide a means of manipulating such values so that they are of a standardised format. This can be achieved in several ways, each involving a change to the data. The easiest approach would be to remove the time value associated with the date value for that field. The change of this data allows for a universal classification for the attribute and makes it far easier to construct the ARFF file afterwards. This process is known as data transformation.

Figure 3: Change of Date

```
elements[elements.ToList().IndexOf(element)] = n.ToString("dd/MM/yyyy");
```

order_no	Extension	Cust_Code	Ship_To	Required_Ship_Date	Sch_Ship_Date	Date_Ship
2023985	0	TSU00200	TSU00201	09/04/2012 00:00	09/03/2012 00:00	#####
2027483	0	TSU00200	TSU00201	11/06/2012 00:00	11/02/2012 00:00	#####
2030828	0	NHSFB500	NHSFB502	01/11/2013 00:00	01/07/2013 00:00	#####
2021609	0	NHSFB500	NHSFB500	09/07/2012 00:00	09/03/2012 00:00	#####
2021609	1	NHSFB500	NHSFB500	09/07/2012 00:00	09/03/2012 00:00	#####
2021231	0	NHSFB500	NHSFB500	08/03/2012 00:00	30/07/2012 00:00	#####
2021231	1	NHSFB500	NHSFB500	08/03/2012 00:00	30/07/2012 00:00	#####
2021231	2	NHSFB500	NHSFB500	17/08/2012 00:00	13/08/2012 00:00	#####
2023605	0	CHR00200	CHR00200	31/08/2012 00:00	24/08/2012 00:00	#####

order_no	Extension	Cust_Code	Ship_To	Required_Ship_Date	Sch_Ship_Date	Date_Ship
2023985	0	TSU00200	TSU00201	09/04/2012	09/03/2012	#####
2027483	0	TSU00200	TSU00201	11/06/2012	11/02/2012	#####
2030828	0	NHSFB500	NHSFB502	01/11/2013	01/07/2013	#####
2021609	0	NHSFB500	NHSFB500	09/07/2012	09/03/2012	#####
2021609	1	NHSFB500	NHSFB500	09/07/2012	09/03/2012	#####
2021231	0	NHSFB500	NHSFB500	08/03/2012	30/07/2012	#####
2021231	1	NHSFB500	NHSFB500	08/03/2012	30/07/2012	#####
2021231	2	NHSFB500	NHSFB500	17/08/2012	13/08/2012	#####
2023605	0	CHR00200	CHR00200	31/08/2012	24/08/2012	#####

Figure 3 outlines the software tool developed to alter the date values stored for each of the related fields. A small element of the code for this utility has been included highlighting the tasks to be performed in order to provide the functionality for the process. The figure illustrates the line of code that performs the change and also the a view of the data before and the resultant format. A full listing of the code used can be viewed within Appendix A.

As a result of this cleaning process the date is provided in a uniform format that the data-mining tool understands and can base all of the dates upon. This makes the classification and clustering process far easier.

Data Removal

In addition to the process of data transformation for those fields that are not uniform, some records may need to be removed as they have no value to the act of data mining. This principle has been noted above however not explained. Within the proposed data set there are some records that have missing values within several records. The lack of information in some key fields can make the record redundant, these key fields include; order_no, cust_code, ordered, shipped, cost, price. This removal of records could have been achieved through the use of the script used for the extraction however these values were substituted to demonstrate this process.

Figure 4 represents a record that falls into this category of a record that should be removed. Note that this record does not contain a value for the customer_code field or the order_no field. The lack of this information means that this record cannot be linked to any customer or order information limiting the relations that can be followed. This type of record is normally created through corruption in the data store or bugs within the software that provides the business interface to the process.

Figure 4: Removal of Records

9981	2018338	0	'MOD'	MODA5301	29/06/2012	25/06/2012	27/06/2012	11/05/2012
9982		0	UNKNWON	UNKNWON	01/06/2012	28/05/2012	09/06/2012	28/05/2012
9983	2019230	0	'EQU'	EQU00101	01/06/2012	28/05/2012	09/06/2012	28/05/2012
9984	2019244	0	'SOU'	SOU00400	22/06/2012	22/06/2012	14/06/2012	28/05/2012

Data-Grouping

In some circumstances when data is taken from real-world applications, there are records that have very similar data within a given field. The similarity of data can make the resulting analysis skewed to some degree. As such similarity exists between the values it may be possible to group such values into what is termed bins. It is important to note however, care is to be taken when attempting grouping of data as the level of granularity is reduced by this. An example of a possible grouping is given in Figure 5. The records displayed contain a location field indicating the location within the organisation that the product resides, it is evident from this depiction that there is a distinct similarity in the values 'OSS-DIRECT' and 'OSS-TBS'.

Figure 5: Grouping

Maureen_Doyle	13	'OSS-TBS'	'OSS'	11/05/2012	1	0
Carolyn_Lloyd	7	'OSS-TBS'	'OSS'	14/05/2012	1	0
Marjo_Plant	18	'OSS-DIRECT'	'OSS'	01/12/2011	78	0
Marjo_Plant	18	'OSS-DIRECT'	'OSS'	01/12/2011	78	78
Marjo_Plant	18	'OSS-DIRECT'	'OSS'	01/12/2011	78	0
Marjo_Plant	18	'OSS-DIRECT'	'OSS'	01/12/2011	78	0

The two values present indicate an area of storage that is not physically separated; this provides a possibility of treating the two values as one group. The two fields could be then grouped as the value 'OSS'. This grouping provides a means for classifying all of the items as being within this location; as a result there is no major skew within the data. In addition to this several other elements have been grouped together dependent on their categorisation into nominal elements. In addition to those values mentioned there are 8 other groups that are similar allowing for classification of the location that stock is retrieved from.

Predictions

As a sales organisation, there are points where the quantity of sales is likely to increase, this may occur at specific periods throughout the year, for example those periods closer to the end of the tax year. There may however be some variation between these times for different customers as they may have differing policies regarding their financing. It is far more likely that an organisation attempt to use up their budget allocated to them at this point within the financial year, in the hopes that this budget be provided to them once again.

Every customer is going to procure office furniture that is specifically tailored to its needs and as such will have a particular range and style in mind when completing such purchases. It is likely that after purchasing one element of a particular range that the organisation would require matching items within this range. This pattern of purchase should be identifiable within the graphical representation of the data by viewing the category. As each product is assigned to a category there are a number of products that may have been purchased, however, only a single

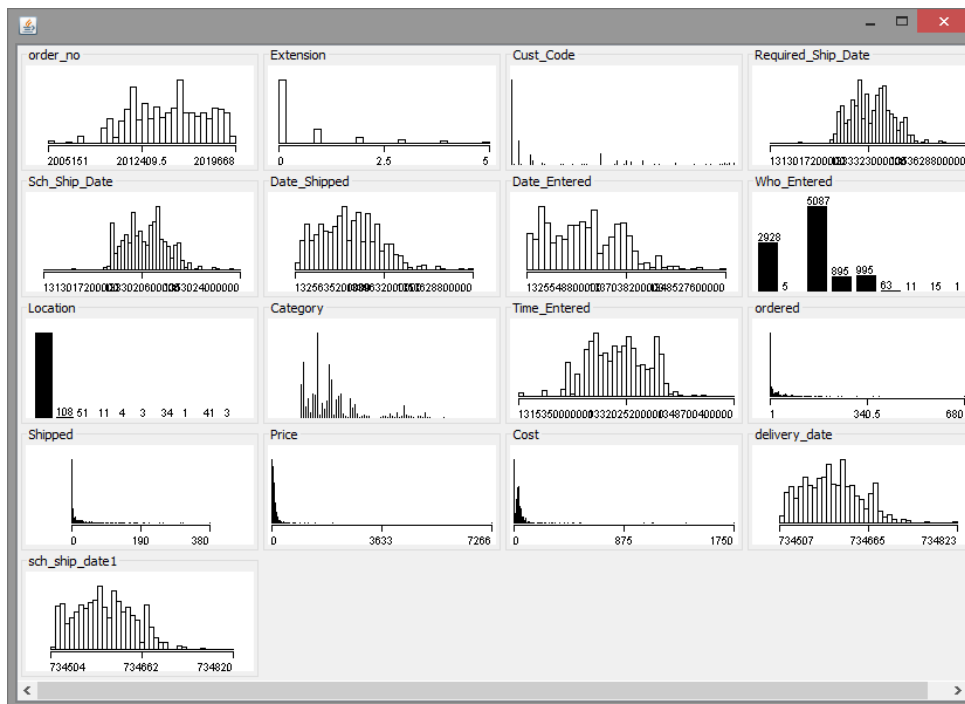
record has been included for the category. This grouping by category generalises the data somewhat however paints an accurate representation of popularity at this abstraction level.

The final area of investigation includes the location from which items are likely to be shipped. Although the items are marginalised into their respective categories, they are also shipped from their respective locations based upon the factory where they are constructed or purchased from. By highlighting the location that serves the greatest demand it provides beneficial information to increase staffing at this location and as such demands can be met.

Pattern Discovery

Following the data preparation stage the dataset should be in a suitable sanitized format so that it may be investigated for possible pattern definition. In undertaking such a process a directed discovery approach will be adhered to providing the key structures and models upon which the overall analyses will be based. The overarching theme of the data-mining process is to identify those customers that purchase specific items at specific times throughout the year with particular focus on recurrent patterns of such purchases as such much of the discovery will be utilised to highlight data eluding to this area. As this goal is tantamount to providing a prediction of further purchase it is necessary to first follow the process of classification of the data.

Figure 6: Initial Load Visualisation



From the initial load of the data into WEKA, as displayed above it is possible to make several observations. Each of the displayed graphs represent the elements contained within each attribute field of the data with their relation to one another. As the order_no field is not required then this may be removed as it has been retained up to this point for reference.

Within the remaining attributes there are several patterns that are apparent, including large spikes within the required_ship_date attribute and the location attribute. This large quantity of elements within the required_ship_date field may provide an insight into a period of high demand. With this pattern and the identification of events within this period it may be consequently possible to determine the cause for such a demand increase.

In addition to demand increases, the disparity between the columns within the location field indicates the vast majority of those items that were shipped have been retrieved from a particular stocking area within the company. Identification of the area may prove useful in ensuring that the stock required to service demand within this area is always maintained.

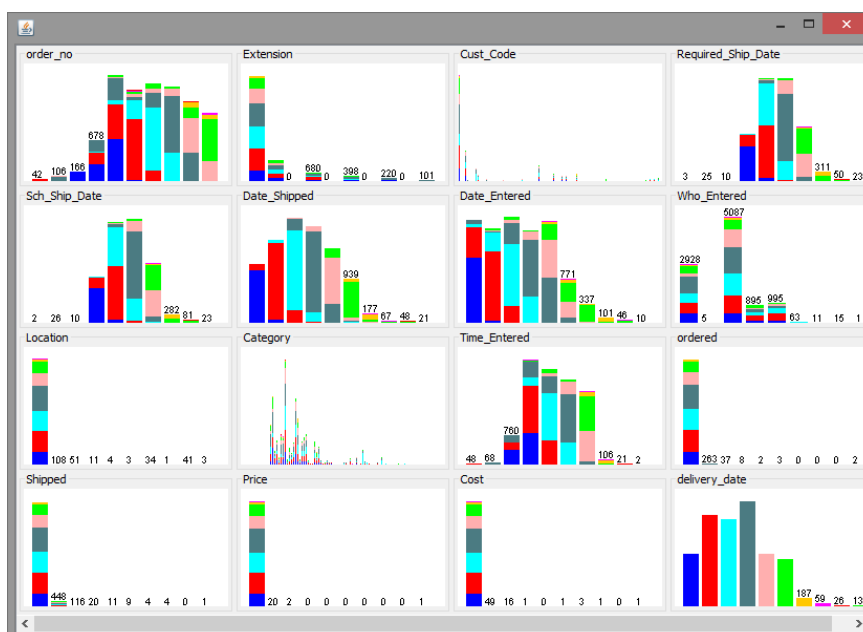
From the graphing information provided it is evident that a large number of items that have been purchased have been from a particular category, eluding to the fact that products from this category may be far more popular than those products from other categories. Categorical information in this context may be derived from the range or the type of product, this makes having such information very useful in determining if one range of products is far more successful than others.

Lastly, from viewing the information provided above it is possible to identify a disparate increase in the number of orders placed by an individual customer. From the graph for customer_code it is possible to note that there is a distinctly higher rate of purchase from one particular customer.

As a result of the interesting observations made from the collected data set, the following sections will outline how to classify, cluster and analyse the findings for three of the four stated points; the location from which products are taken, the increase in demand for the time period and the products from the specific category. It is hoped that the analyses conducted upon this data will elucidate the reasons behind such dramatic changes in the visualisation.

Following this initial load into WEKA a discretisation algorithm was used in order to ensure that all values had been rendered discrete and did not cause the categorisation of the elements within the dataset to fail.

Figure 7: Post-Discretisation View



Although the data has changed marginally, the patterns observed still follow the same general shape and structure providing for the same view of the data as was present prior to the discretisation process.

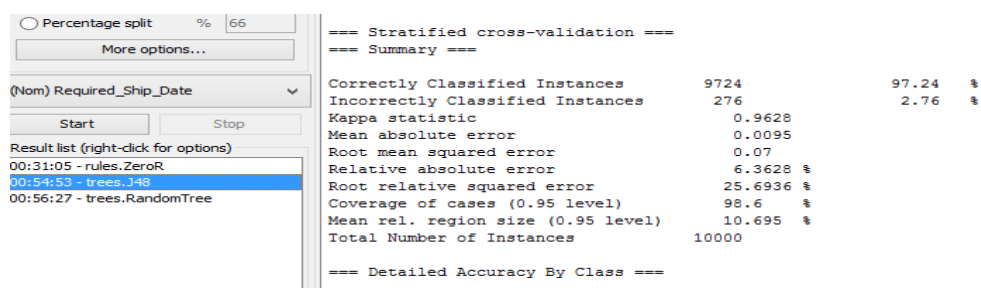
Classification

As outlined in a previous section, the process of classification allows for the data analyst to provide a prediction of future events based on the similarity of attributes of previous events. Although this prediction can be made it is not guaranteed to be one hundred percent accurate it is merely predictive. The purpose of this classification is to construct a training model that may be applied to testing data sets in order to represent meaningful predictive information to the user. (7)

In an attempt to provide an adequate classification model for the data set used, several were considered and tested against a subset of the data. This process led to the disregarding of the ZeroR algorithm, the ZeroOne algorithm and the RandomTree algorithm.

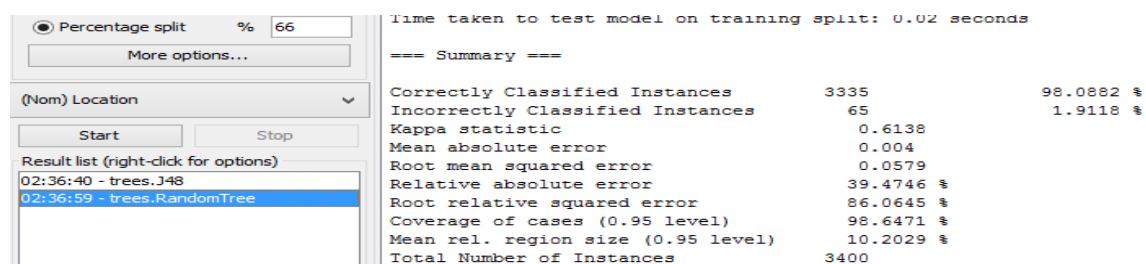
The approach opted for in lieu of those previously mentioned was the utilisation of the J48 algorithm. This method was selected primarily because it appeared to best fit the data being utilised and consequently had lower inaccurate classifications based upon the model. The J48 algorithm classification was utilised with relation to the scheduled_ship_date attribute. The relative and absolute error values have been depicted within figure 8.

Figure 8: Scheduled_Ship_Date Classification Error



In contrast to this, a different algorithm was used to classify the stock location items. The algorithm chosen for this purpose was again based upon the best possible coverage and correctness of the element classification. This correct classification ratio ensures that the clustering and consequent analyses based upon the results that are achieved from the employ of such models are reliable and as accurate as possible. The chosen method of classification of the location attribute is manifested in the form of the RandomTree algorithm. A depiction of this classification training set has been outlined below.

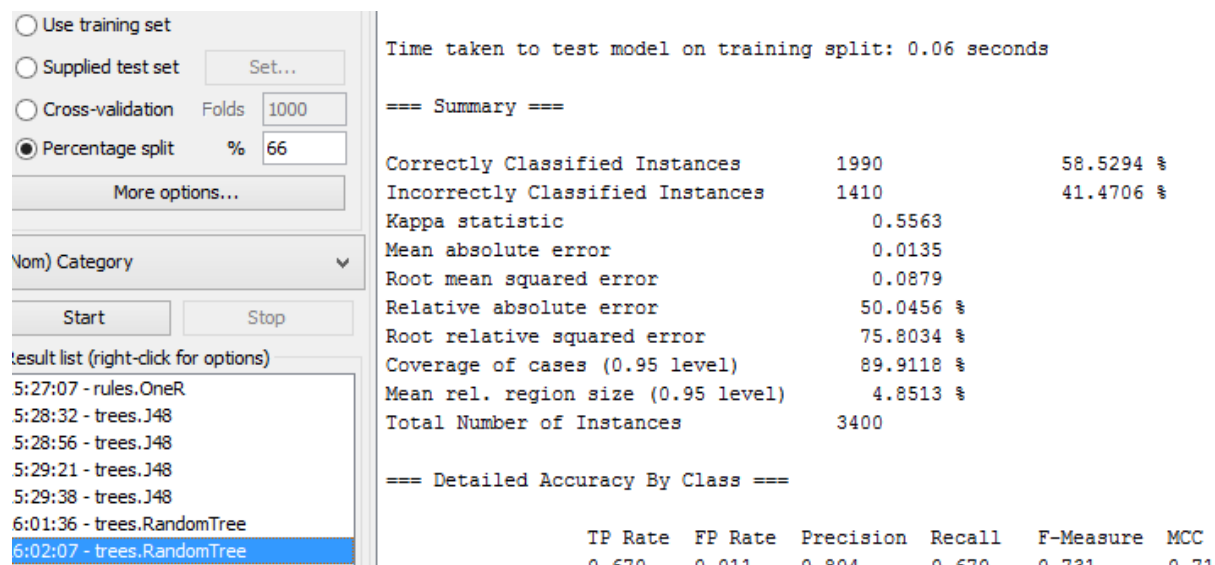
Figure 9: Location RandomTree Error



Based upon this verbose description of the model classification, it is sensible to assume that this is most likely the best possible algorithm for the classification of this attribute, given the error percentage and the correct classification instances.

The third area of interest includes the perusal of data pertaining to the category to which a product belongs. As mentioned in a previous section, this information may elucidate as to which products may continuously be in high demand. In order to provide a classification model for this element an entirely different approach had to be used as there are a vast number of nominal values present. The categories utilised were further grouped in an attempt to make it easier to classify. As the values contained within this attribute are all nominal the classification percentage is a lot lower than those as provided for numeric or date values. The algorithm used for this classification is the RandomTree algorithm, however alterations were made to its initialisation values.

Figure 10: Category Random Tree Classification



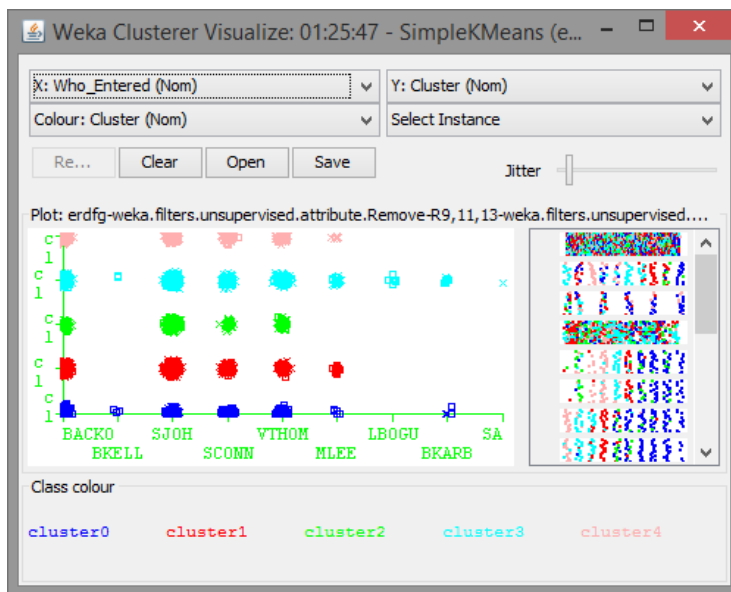
All of the classification models constructed within this stage of the process have been saved and retained for future testing of the data sets. Following the classification it is necessary to perform further categorisation in the form of clustering to identify key relationships and patterns within the data. These models have been utilised for training the methods utilised for the testing process and provides for a far more accurate representation of the patterns within the data. For each of the classifications identified a 66% split has been utilised between training data and for testing data, utilising 66% of the dataset for training and 34 for testing or re-evaluating the constructed model.

Clustering

In addition to the process of classification to construct a base model, some of the data has also been grouped in order to extract meaningful pattern analysis from it. For this very reason, some method of clustering is applied to the data following the construction of a classification model. This process of clustering makes it far easier for the observer to identify areas of patterns and categorise them as such. (8)

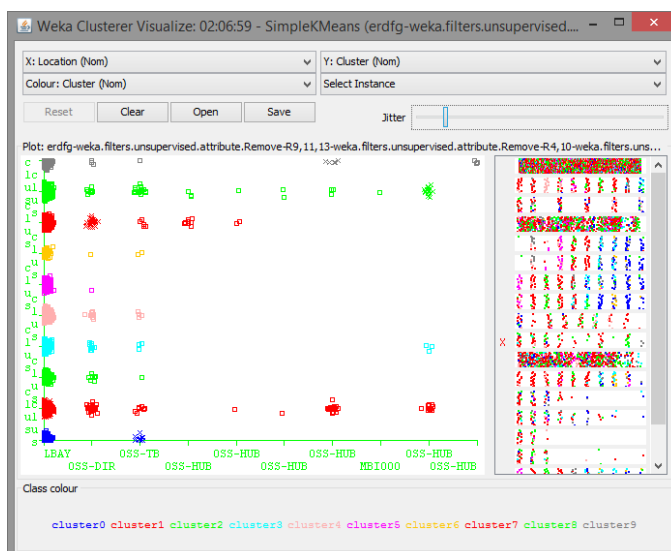
Clustering is a process used to group data elements together into an area that encompasses relative or similar elements. The purpose of this form of model creation allows for outliers to be seen far more easily, any data that does not fall within the trained clusters will become far more obvious to the observer when employing this form of modelling. For completeness in this instance clustering has been performed on the data set to satisfy each of the three observation areas, highlighting the suitability of this approach in comparison with the use of classification and the use of the associated algorithms.

Figure 11: Scheduled_Ship_Date KMeans



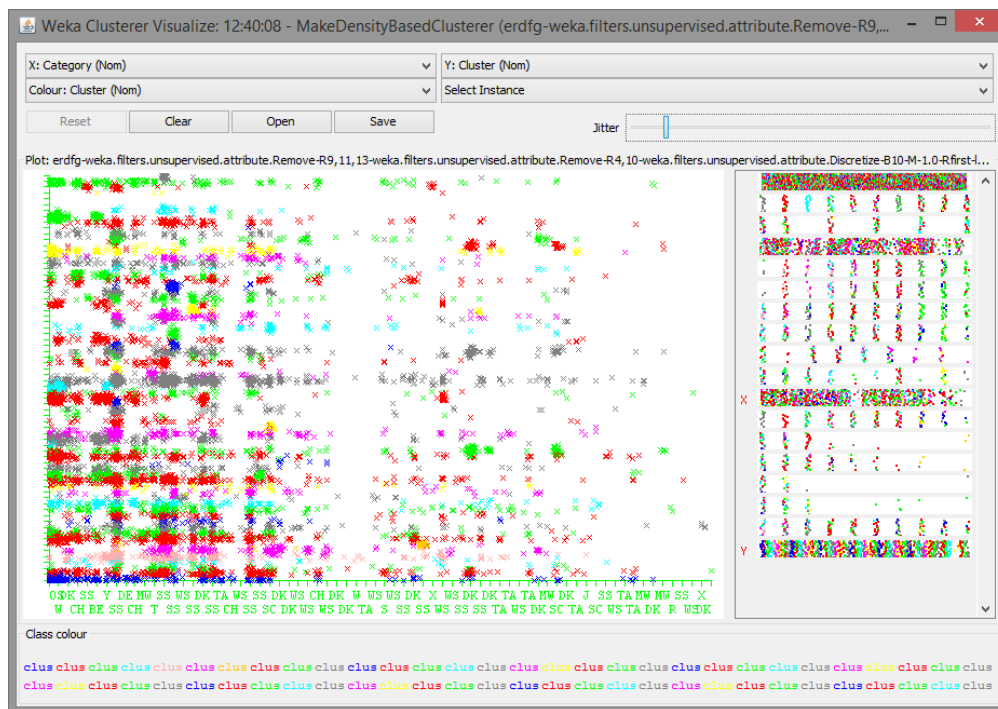
When attempting to model the data for the scheduled_ship_date with the clustering algorithm KMeans, the following representation was achieved. This depiction of the clustering of scheduled_ship_date allows for the viewing of items clustered by the date that they have been scheduled to ship alongside the employees who had entered such information. It is evident from this representation that this may not be the best possible way in which to classify or represent such data, for this particular data-mining request it would be more beneficial to investigate through the use of time-series analysis.

Figure 12: Location Clustering Representation



This representation is similar to that of the scheduled_ship_date in that the use of clustering to represent such patterns within the data is insufficient due to the grouping provided for such nominal values. The diagram below shows clustering as performed upon the category of the items within the dataset.

Figure 13: Category Clustering



This clustering approach utilised for the category has some merit and allows for visualisation of the products that are related to each category and that come from particular areas within stock. In this case the MakeDensityBasedClusterer clustering algorithm was utilised. The use of clustering here, better highlights the definitions between the categories and as such makes it far more clear which categories are more popular than others.

Association

One key area of data mining has not been covered within context of this example, this approach is known as associative machine learning. Associative learning is concerned with identifying those elements that are recurring in conjunction with other elements and is particularly useful when attempting to determine what products customers may purchase in complement to the items that they actually required. (9)

Although not covered within the examples provided above, this form of mining is possible as the dataset contains data relating to the category within which the purchase product lies. This data has been used above to determine which categories are popular at what intervals, however could be used for other purposes inclusive of investigating which product categories are purchased in conjunction with other categories and the frequency with which this procurement occurs.

There are three main algorithms pertaining to associative learning; FPGrowth, FilteredAssociation and the much esteemed Apriori algorithm. These algorithms work based upon the notion of the frequency with which elements appear in conjunction with one another and as such may provide the observer with a degree of certainty that a customer may purchase a group of items based on environmental or personal circumstances.

In this case, the Apriori algorithm could be applied to the category field of the data set providing information as to the categories that relate to one another. This would be beneficial to

attempt upon further processing of the data by the company providing them with possibly new marketing options. To provide the data in a format that allows this type of mining would require a different approach to the pre-processing of the data-set however and would not conform to the data in the approaches mentioned in previous sections.

Secondary Pre-Processing

In the case of the data relating to the category attribute, much of the data is skewed because of some of the other attributes that aren't required. As is the case, some secondary pre-processing is required in order to provide a more accurate result set upon which to base the final analysis of this area.

Attribute Removal

To improve the accuracy of the classification of the category attribute, several attributes were removed including the line_no attribute, the extension attribute, ordered, shipped, price and cost. These values are not relevant to the information that is to be analysed within the confines of this text. As such the removal of these elements will be achieved at this stage.

Figure 14: Secondary Attribute Removal

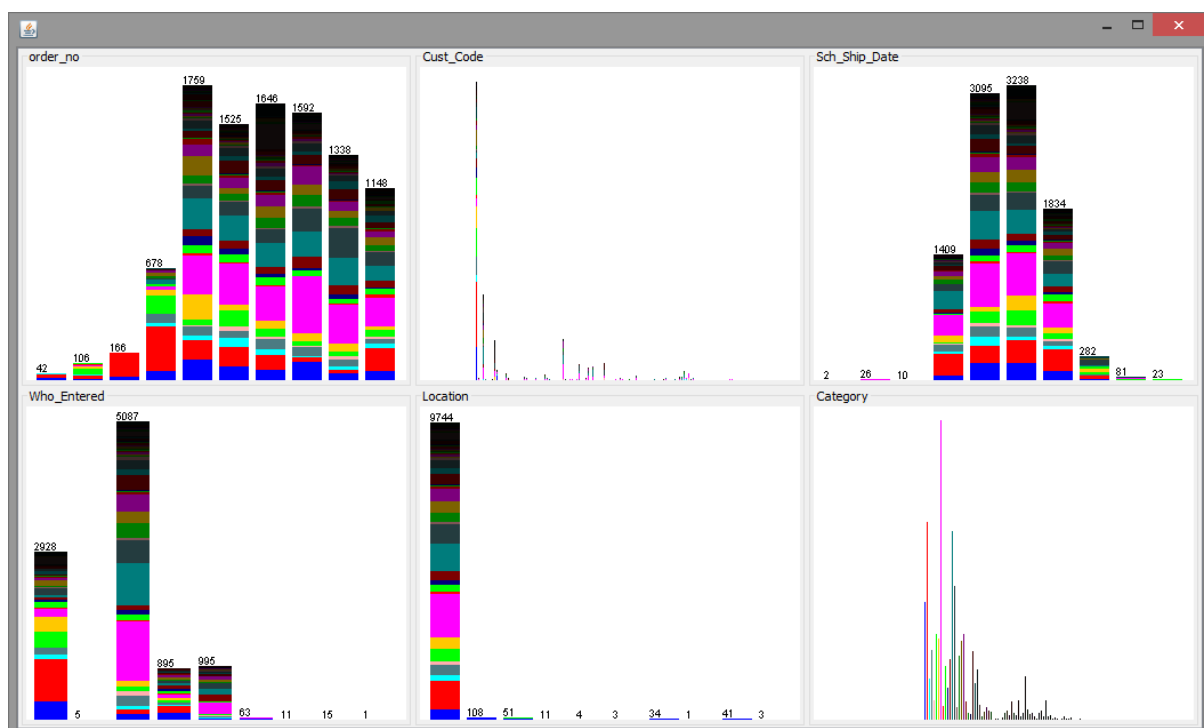


Figure 14, represents the change in data after the removal of such records. It is evident from the graphic that there is a vast difference in the visualisation of the data patterns when compared against those that are present within figures 6 and 7.

The reduced number of attributes allows for the classifications to be made more accurate providing cleaner information to them and as such making it far easier to group nominal elements as opposed to the classification of numerical elements.

Analyses and Findings

From the classification performed upon the `scheduled_ship_date` and the graphical representations provided, it is possible to deduce that there was a radical increase in the number of orders that were required to be delivered approximately January through to March. This appears to be the year end for most companies and complies with the proposed predictions. This data alongside personal knowledge of the operation of the company and its clients could be a direct result of end of fiscal year for most companies. Although this prediction is not guaranteed, the purpose of classification is to allow the prediction of future rises or falls in demands and as such fulfils this requirement in this context. The possession of this knowledge allows the company to increase available stock to meet demand at this point.

With regards to the `randomtree` classification applied to the location information relating to orders within the data set there are a number of apparent details that may be discussed. The most obvious factor that can be observed from the graphical representations is that the majority of orders ship their products from the loading bay "LBAY" whereas only a small amount of orders are shipped from other locations. When offsetting this information against the category code of products, it is possible to identify those product categories that are likely to be procured from other companies.

The majority of items that are not shipped from the location labelled "LBAY" are for the most part purchased as complete units from suppliers, however those items prefixed with the three letter code "OSS" are items that are constructed at a factory owned by the company and are shipped directly from the location they are created.

The identification of category codes was a complex task as the field contained several nominal values which are far harder to classify when being used in a data set that is comprised of a mix of numeric and nominal values, the resultant classification verbosity provided above is testament to the possibility and or probability of incorrect classification within the dataset.

When plotting the category against customer code on an axis, it is possible to determine the product categories that are most purchased by each given customer and as such it is possible to target marketing companies for said products at these customers in future. A further benefit to the presence of this information is that it is possible to determine an increase in demand for certain product categories based on the time at which they have been entered into the system.

Reflection

The data set provided for this scenario, serves a multitude of purposes however was not provided in a usable format. This required much preprocessing in order for it to be accepted by the data mining tool. After pre-processing the data and removing several of the attributes, the data was represented in some form of sanitised state and as such could be understood and accurately represented by the data mining tool.

The data mining tools (WEKA) itself had several stringent controls on the formats in which the data may be supplied to it. This restriction on the data format caused several issues upon the initial loading as many of the nominal attributes within the dataset contained spaces. Following this initial attempt to load in the data, further pre-processing was administered and as such the spaces within the fields were replaced with an underscore character. In addition to the formatting issues with the inclusion of spaces, date formats were far more difficult to handle and as a result the code provided within appendix A was constructed to standardise.

Although not all of the predictions that were made were proved, the vast majority of those assumptions made were proved by the patterns visible within the graphs. The increase in the scheduled_ship_date is supported by the graphical representation and thus proving that this increase occurs at certain points throughout the year.

To increase the beneficial information retrieved from this dataset, an associative approach may be applied to this data. The apriori algorithm would be applied to the category attribute and as such those categories that are associated with one another would become obvious.

References

1. **Mohammed J. Zaki, Wagner Meira, Jr.** *Data Mining and Analysis: Fundamental Concepts and Algorithms*. s.l. : Cambridge Press, 2014. p. 9.
2. **Jiawei Han, Micheline Kamber, Jian Pei.** *Data Mining: Concepts and Techniques: Concepts and Techniques*. s.l. : Elsevier, 2011. p. 3.
3. **Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman.** *Mining of Massive Datasets*. s.l. : Cambridge University, 2014. p. 4.
4. **KCStores.** *Sales Data Set*. s.l. : KCStores, 2012-2014.
5. **Jiawei Han, Micheline Kamber, Jian Pei.** *Data Mining, Southeast Asia Edition: Concepts and Techniques*. s.l. : Morgan Kaufmann, 2006. pp. 6-8.
6. **Ian H. Witten, Eibe Frank, Mark A. Hall.** *Data Mining: Practical Machine Learning Tools and Techniques*. s.l. : Elsevier, 2001. pp. 10-15.
7. **Aggarwal, Charu C.** *Data Classification: Algorithms and Applications*. s.l. : CRC Publishing, 2014. pp. 2-9.
8. **Charu C. Aggarwal, Chandan K. Reddy.** *Data Clustering: Algorithms and Applications*. s.l. : CRC Publishing. pp. 4-6.
9. **Jiawei Han, Micheline Kamber, Jian Pei.** *Data Mining, Southeast Asia Edition: Concepts and Techniques*. s.l. : Morgan Kauffmann, 2009. pp. 4-8.

Appendix A

Extraction Script

```
select distinct
o.order_no,
isnull(o.ext,0) Extension,
'''+isnull(substring(upper(o.cust_code),0,4),'UNKNOWN')+'''' Cust_Code,
isnull(o.ship_to,'UNKNOWN') Ship_To,
cast(isnull(o.req_ship_date,'9999-12-31')as date) Required_Ship_Date,
cast(isnull(o.sch_ship_date,'9999-12-31 23:59:59')as date) Sch_Ship_Date,
cast(isnull(o.date_shipped,'9999-12-31 23:59:59')as date) Date_Shipped,
cast(isnull(o.date_entered,'9999-12-31 23:59:59')as date) Date_Entered,
coalesce(nullif(replace(replace(o.cust_po,'',''),' ','_'),),'UNKNOWN') Cust_PO,
'''''+isnull(upper(o.who_entered),'sa')+'''' Who_Entered,
isnull(o.status,'V') Status,
coalesce(nullif(REPLACE(replace(o.phone,'',''),' ','_'),),'UNKNOWN') Phone,
isnull(ol.line_no,1) Line_no,
'''''+isnull(upper(ol.location),'UNKNOWN')+'''' Location,
'''''+isnull(upper(m.category),'UNKNOWN')+'''' Category_Code,
cast(isnull(ol.time_entered,'9999-12-31 23:59:59 000')as date) Time_Entered,
ol.ordered,
isnull(ol.shipped,0) Shipped,
isnull(ol.price,0) Price,
isnull(ol.cost,0) Cost,
ol.delivery_date,
ol.sch_ship_date

from orders o
join ord_list ol on ol.order_no=o.order_no
LEFT JOIN inv_master m on m.part_no=ol.part_no
where date_entered BETWEEN '01 Jan 2012' AND '31 Dec 2014'and
isnull(date_entered,'9999-12-31 23:59:59') BETWEEN '01 Jan 2012' AND '31 Dec 2014'
AND ol.ordered>0
AND ol.delivery_date IS NOT NULL
AND date_shipped IS NOT NULL
AND coalesce(nullif(o.phone,''),'UNKNOWN')!='UNKNOWN'
AND coalesce(nullif(o.cust_po,''),'UNKNOWN')!='UNKNOWN'
```

Pre-Processing Date Formatting

```
public static void DateFormatting()
{
    foreach (string line in lines)
    {
        string[] elements = line.Replace("\r", "").Split(new char[] { ',' });
        foreach (string element in elements)
        {
            if (!string.IsNullOrEmpty(element))
            {
                DateTime n = new DateTime();
                if (DateTime.TryParse(element, out n))
                {
                    elements[elements.ToList().IndexOf(element)] = n.ToString("dd/MM/yyyy");
                }
            }
        }
        lines2.Add(String.Join(",", elements).Replace("\n", ""));
    }
}
```