Question 1)

1. Perspective Analytics
2. **Kafka**

* Apache™ Kafka is a fast, scalable, durable, and fault-tolerant publish-subscribe messaging system. Kafka is often used in place of traditional message brokers like JMS and AMQP because of its higher throughput, reliability and replication.
* Kafka works in combination with Storm, HBase and Spark for real-time analysis and rendering of streaming data. Kafka can message geospatial data from a fleet of long-haul trucks or sensor data from heating and cooling equipment in office buildings. Whatever the industry or use case, Kafka brokers massive message streams for low-latency analysis in Enterprise Apache Hadoop.
* Kafka’s system design can be thought of as that of a distributed commit log, where incoming data is written sequentially to disk. There are four main components involved in moving data in and out of Kafka:
  + Topics
  + Producers
  + Consumers
  + Brokers
* Topic is a user-defined category to which messages are published. Kafka Producers publish messages to one or more topics and Consumers subscribe to topics and process the published messages. Finally, a Kafka cluster consists of one or more servers, called Brokers that manage the persistence and replication of message data (i.e. the commit log).
* Reference: http://hortonworks.com/apache/kafka/#section\_2

1. **Apache Spark**

* Spark is an open source big data processing framework built with the intention of Speed, easy-to-use and sophisticated analytics. First developed by UC Berkley and then made open source by Apache.
* It is a flexible in-memory framework which provides big data processing on diverse types of datasets (text dataset or Graph dataset etc.) and also on the source of data (real-time or batch processing).
* It’s APIs in Java, Scala & Python are built in such a way that it provides good set of operations via SQL queries, stream-processing, Graph analysis and Machine-learning libraries.
* Spark holds intermediate results in memory rather than writing them to disk which is very useful especially when you need to work on the same dataset multiple times. It’s designed to be an execution engine that works both in-memory and on-disk. Spark operators perform external operations when data does not fit in memory. Spark can be used for processing datasets that larger than the aggregate memory in a cluster.
* Spark will attempt to store as much as data in memory and then will spill to disk. It can store part of a data set in memory and the remaining data on the disk. You have to look at your data and use cases to assess the memory requirements. With this in-memory data storage, Spark comes with performance advantage.
* It Supports more than just Map and Reduce functions, optimizes arbitrary operator graphs, lazy evaluation of big data queries which helps with the optimization of the overall data processing workflow.
* Reference: <https://www.infoq.com/articles/apache-spark-introduction>

1. **Big Data**

* Big data is a term being used off lately and is the new hot stuff in the tech industry. There have been multiple attempts to define big data but there is no proper definition.
* Basically, any data which is big enough to be processed by traditional technological solutions is called big data. These kind data require sophisticated technologies and frameworks to make sense. Big Data is generally referred by 3 V’s :
  + ***Volume***: We currently see the exponential growth in the data storage as the data is now more than text data. We can find data in the format of videos, music and large images on our social media channels. It is very common to have Terabytes and Petabytes of the storage system for enterprises. As the database grows the applications and architecture built to support the data needs to be reevaluated quite often.
  + ***Variety***: Data can be stored in multiple format like database, excel, csv, access or for the matter of the fact, it can be stored in a simple text file. Sometimes the data is not even in the traditional format as we assume, it may be in the form of video, SMS, pdf or something we might have not thought about it. It is the need of the organization to arrange it and make it meaningful. It will be easy to do so if we have data in the same format, however it is not the case most of the time.
  + ***Velocity***: The data growth and social media explosion have changed how we look at the data. There was a time when we used to believe that data of yesterday is recent. The matter of the fact newspapers is still following that logic. However, news channels and radios have changed how fast we receive the news. Today, people reply on social media to update them with the latest happening. On social media sometimes a few seconds old messages (a tweet, status updates etc.) is not something that interests users. They often discard old messages and pay attention to recent updates. The data movement is now almost real time and the update window has reduced to fractions of the seconds.
* In 2008, here was an outage at Netflix due to which many customers were left in the dark. While some could still access the streaming services, most of them could not. Some customers managed to get their rented DVDs whereas others failed. The outage made the management think about the possible future problems and the hence, it turned to Big Data. It analyzed high traffic areas, susceptible points, and network throughput etc. using that data and worked on it to lower the downtime if a future problem arises as it went global.
* Reference: <http://www.thewindowsclub.com/what-is-big-data>

<http://www.sas.com/en_us/insights/big-data/what-is-big-data.html>

1. **Trust Based Recommender Systems**

* Trust based Recommender Systems have become an integral part of web-based systems. Many e-commerce systems, online-streaming services etc. have been using recommender systems to service the best to their customers.
* The way these systems are designed are taken from the real-world scenario.
  + Consider a scenario a situation some decades back where there was no use of technology and a person wants to buy a product. So he/she will ask their friends about the product. If they don’t know, they will ask their friends. Based on collective input received by the person, he will determine to buy a product or not.
  + Why to ask a friend? This is because that person has trust in his friends and this is how based on trust a person decides to buy a product or not.
* Similarly, trust based recommender systems work in such a way that if you are browsing through Amazon searching for a product, Amazon’s recommender system will check your profile (previous search, social media profiling etc). Based on that profile, amazon will try to find similar people and check what products they have brought and will recommend you to buy those products.
* Netflix also uses a recommender system which will recommend you movies, series etc. based on what similar people to you have been watching.
* Reference: <http://research.microsoft.com/en-us/um/people/borgs/papers/trust.pdf>

1. **Biologically inspired Data Mining**
2. **Knowledge Discovery**

* In this digital age where every click or action is generating data, the amount of data getting generated is quite intimidating. To give a glimpse of how much data is getting generated, these are the facts:
  + Europe's Very Long Baseline Interferometry (VLBI) has 16 telescopes, each of which produces 1 Gigabit/second of astronomical data over a 25-day observation session
  + Walmart reported to have 24 Tera-byte DB
  + Computer Speed doubles every 18 months
  + Total storage doubles every 9 months
* So the science of making sense of all the data, finding patterns in it is called Knowledge Discovery. This data can consist of anything from boolean values, sports scores, web visit statistics, laboratory analyses, or other scientific measurement.
* Under Knowledge discovery we have following parts:
  + Data Cleaning
  + Data Integration
  + Data Selection
  + Data Transformation
  + Data Mining
  + Pattern Evaluation
  + Knowledge Presentation
* Reference: <http://www.semagix.com/knowledge-discovery.htm>

<https://www.tutorialspoint.com/data_mining/dm_knowledge_discovery.htm>

1. **Class Label**

* The term class label is usually used in the context of supervised machine learning, and in data classification in particular, where one is given a set of examples of the form (attribute values, classLabel) and the goal is to learn a rule that computes the label from the attribute values. The class label always takes on a finite number of different values.
* For example, if we want to find whether a person will become homeless or not. We have a data set consisting of fields (origin, education, salary, isHomeless). So here origin, education and salary are the attributes and isHomeless is the class label. So the training and testing set will contain this class label which will help us to create a model but the actual data will not have this field.

1. **Predictive Analytics**

* It is a branch of advanced analytics which is used to make future predictions about some event. It uses many techniques such as data mining, statistics, machine learning, AI etc. to predict the future. The pattern found in historical data or transactional data is used to derive the predictions.
* Predictive models use known results to develop (or train) a model that can be used to predict values for different or new data. Modeling provides results in the form of predictions that represent a probability of the target variable (for example, revenue) based on estimated significance from a set of input variables.
* Predictive analytics has been in the world for decades but it has gathered heat over the past couple of years. Many organizations are turning towards predictive analytics to increase their bottom line and gain competitive advantage.
* With interactive and easy-to-use software becoming more prevalent, predictive analytics is no longer just the domain of mathematicians and statisticians. Business analysts and line-of-business experts are using these technologies as well.
* Organizations have been using predictive analytics for many things like Fraud Detection, optimizing market campaigns, improve operations, reduce risk etc.

1. **Analytic**

* Analytics is an encompassing and multidimensional field that uses mathematics, statistics, predictive modeling and machine-learning techniques to find meaningful patterns and knowledge in recorded data.
* Today, we add powerful computers to the mix for storing increasing amounts of data and running sophisticated software algorithms – producing the fast insights needed to make fact-based decisions. By putting the science of numbers, data and analytical discovery to work, we can find out if what we think or believe is really true. And produce answers to questions we never thought to ask. That’s the power of analytics.
* We are not only generating vastly more data but our ability to harness and analyse this data has improved massively over recent years. We can now analyse large volumes of fast moving data from different data sources to gain insights that were never possible before. Different types of analytics approaches allow us to analyse numbers, text, photos and even voice and video sequences.
* Analytics has been in practice in many fields, like:
  + Sports: In tennis we use a system called SlamTracker that records player performance providing real-time statistics and comprehensive match analytics.
  + Healthcare: To predict the likelihood of certain disease happening.
  + Love: Companies like Tinder, eharmony etc. have created a prediction model which takes into account the different variable relating to personality traits, behaviors & social skills and helps find a perfect match.
* Reference: <https://www.linkedin.com/pulse/20130624053353-64875646-what-the-hell-is-analytics>

[http://www.sas.com/en\_us/insights/analytics/what-is-analytics.html#](http://www.sas.com/en_us/insights/analytics/what-is-analytics.html)

1. **Hadoop**

* Hadoop is an open-source, Java-based programming framework the supports the storage and processing of large datasets in a distributed environment.
* Hadoop turned out to be very important for many reasons, like:
  + Ability to store and process large amounts of data very quickly.
  + Being distributed in nature, the computing power is also very fast.
  + Data and processes are protected against hardware failure. If one of the nodes fail, then processes are redirected to another node so that computing does not stop and multiple copies of data are kept.
  + Flexibility is one of the key features. Unlike traditional relational databases, no data pre-processing is required. Any type of data can be stored without any problem.
  + Scalability is one big advantage.
* Reference: <http://www.sas.com/en_us/insights/big-data/hadoop.html>

1. **Deep Belief Network**

* Deep belief Network, also known as deep neural network is a generative graphical model. It consists of multiple layers of latent variables (which has hidden units) where there are connections between the layers but not between the units within the latent layers.
* If a DBN is trained in an unsupervised way, then it can learn to probabilistically reconstruct all the inputs which in turn acts as feature detectors for the inputs. After this, the DBN can be further trained in a supervised way to perform classification.
* DBN can be viewed as a composition of simple, unsupervised networks such as restricted Boltzmann machines (RBMs) or auto-encoders, where each sub-network's hidden layer serves as the visible layer for the next. This also leads to a fast, layer-by-layer unsupervised training procedure, where contrastive divergence is applied to each sub-network in turn, starting from the "lowest" pair of layers.
* Reference: <https://en.wikipedia.org/wiki/Deep_belief_network>

1. **Machine Learning**

* Machine learning is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look.
* The process of machine learning is similar to that of [data mining](http://searchsqlserver.techtarget.com/definition/data-mining). Both systems search through data to look for patterns. However, instead of extracting data for human comprehension -- as is the case in data mining applications -- machine learning uses that data to detect patterns in data and adjust program actions accordingly.
* Machine learning algorithms are often categorized as being supervised or unsupervised. Supervised algorithms can apply what has been learned in the past to new data. Unsupervised algorithms can draw inferences from datasets.
* There are many examples of machine learning problems, like:
  + optical character recognition: categorize images of handwritten characters by the letters represented
  + face detection: find faces in images (or indicate if a face is present)
  + spam filtering: identify email messages as spam or non-spam
  + topic spotting: categorize news articles (say) as to whether they are about politics, sports, entertainment, etc.
* Reference: <http://www.cs.princeton.edu/courses/archive/spr08/cos511/scribe_notes/0204.pdf>

<http://www.sas.com/en_id/insights/analytics/machine-learning.html>

1. **Feature Selection**

* In this age of huge data inflow, not every part of data is useful for any given problem. Suppose, we have a data-mining problem and the data set consists of 40 fields. Not all the fields will help in mining the data.
* Basically, selection of features or attributes which are more relevant to the analytic problem is called feature selection. So in short we create a subset of the data we have to use for model construction.
* Main advantage of feature selection is that it helps in removing unneeded, irrelevant and redundant attributes from data that do not contribute to the accuracy in the model and in fact may decrease the accuracy.
* Fewer attributes is always better as it will help reduce the complexity and make it easier to comprehend the model.
* There are three general classes of feature selection algorithms:
  + Filter Methods
  + Wrapper Methods
  + Embedded Methods
* Reference: <http://machinelearningmastery.com/an-introduction-to-feature-selection/>

1. **Business Intelligence**

* Business intelligence (BI) refers to a variety of software applications used to analyze an organization’s raw data. BI as a discipline is made up of several related activities, including data mining, online analytical processing, querying and reporting.
* The potential benefits of business intelligence programs include accelerating and improving decision making; optimizing internal business processes; increasing operational efficiency; driving new revenues; and gaining competitive advantages over business rivals. BI systems can also help companies identify market trends and spot business problems that need to be addressed.
* BI technology also includes data visualization software for designing charts and other infographics, as well as tools for building BI dashboards and performance scorecards that display visualized data on business metrics and key performance indicators in an easy-to-grasp way.

1. **Data Warehouse**

* A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but it can include data from other sources. It separates analysis workload from transaction workload and enables an organization to consolidate data from several sources.
* In addition to a relational database, a data warehouse environment includes an extraction, transportation, transformation, and loading (ETL) solution, an online analytical processing (OLAP) engine, client analysis tools, and other applications that manage the process of gathering data and delivering it to business users.
* There are two approaches to creating a Data Warehouse:
  + **Bottom-up approach**: In the bottom-up design approach, the data marts are created first to provide reporting capability. A data mart addresses a single business area such as sales, Finance etc. These data marts are then integrated to build a complete data warehouse.  The integration of data marts is implemented using data warehouse bus architecture.
  + **Top-down approach:** In the top-down design approach the, data warehouse is built first. The data marts are then created from the data warehouse.
* **Reference:** <https://docs.oracle.com/cd/B10500_01/server.920/a96520/concept.htm>

<http://searchsqlserver.techtarget.com/definition/data-warehouse>

1. **Cross-Validation**

* Cross-validation is a more sophisticated holdout training and testing procedure. We would like not only a simple estimate of the generalization performance, but also some statistics on the estimated performance, such as the mean and variance, so that we can understand how the performance is expected to vary across datasets. This variance is critical for assessing confidence in the performance estimate.
* Cross-validation also makes better use of a limited dataset. Unlike splitting the data into one training and one holdout set, cross-validation computes its estimates over all the data by performing multiple splits and systematically swapping out samples for testing.
* The data set is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the test set and the otherk-1 subsets are put together to form a training set.
* Then the average error across all k trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set k-1 times.
* The variance of the resulting estimate is reduced as k is increased. The disadvantage of this method is that the training algorithm has to be rerun from scratch k times, which means it takes k times as much computation to make an evaluation.
* A variant of this method is to randomly divide the data into a test and training set k different times. The advantage of doing this is that you can independently choose how large each test set is and how many trials you average over.
* Reference: <https://www.cs.cmu.edu/~schneide/tut5/node42.html>

<http://scikit-learn.org/stable/modules/cross_validation.html>

1. **Graph Database**

* A graph database, also called a graph-oriented database, is a type of NoSQL database that uses graph theory to store, map and query relationships.
* A graph database is essentially a collection of nodes and edges. Each node represents an entity (such as a person or business) and each edge represents a connection or relationship between two nodes. Every node in a graph database is defined by a unique identifier, a set of outgoing edges and/or incoming edges and a set of properties expressed as key/value pairs. Each edge is defined by a unique identifier, a starting-place and/or ending-place node and a set of properties.
* Graph databases are well-suited for analyzing interconnections, which is why there has been a lot of interest in using graph databases to mine data from social media. Graph databases are also useful for working with data in business disciplines that involve complex relationships and dynamic schema, such as supply chain management, identifying the source of an IP telephony issue and creating "customers who bought this also looked at..." recommendations.
* Reference: <https://neo4j.com/developer/graph-database/>

1. **Confusion Matrix**

* A confusion contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix.
* A confusion matrix looks like the below:

|  |  |  |
| --- | --- | --- |
|  | True - no | True - yes |
| Pred - no | TN | FN |
| Pred - yes | FP | TP |

* Different components of the confusion matrix are:
  + true positives (TP): These are cases in which we predicted yes, and actual class is also true.
  + true negatives (TN): We predicted no, and the actual class is no.
  + false positives (FP): We predicted yes, and the actual class is no.
  + false negatives (FN): We predicted no, but the actual class is yes.
* Reference: <http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/>

1. **Split Validation**

* This operator performs a simple validation i.e. randomly splits up the dataset into a training set and test set and evaluates the model. This operator performs a split validation in order to estimate the performance of a learning operator. It is mainly used to estimate how accurately a model will perform in practice.

1. **Sentiment Analysis**

* Sentiment analysis, otherwise known as opinion mining is the process of determining the emotional tone behind a series of words, used to gain an understanding of the attitudes, opinions and emotions expressed within an online mention.
* The applications of sentiment analysis are broad and powerful. The ability to extract insights from social data is a practice that is being widely adopted by organizations across the world.
* Shifts in sentiment on social media have been shown to correlate with shifts in the stock market.
* The Obama administration used sentiment analysis to gauge public opinion to policy announcements and campaign messages ahead of 2012 presidential election.
* Reference: <https://en.wikipedia.org/wiki/Sentiment_analysis>

<https://www.lexalytics.com/technology/sentiment>

1. **Feature Extraction**

* Feature extraction a type of dimensionality reduction that efficiently represents interesting parts of an image as a compact feature vector. This approach is useful when image sizes are large and a reduced feature representation is required to quickly complete tasks such as image matching and retrieval.
* Feature detection, feature extraction, and matching are often combined to solve common computer vision problems such as object detection and recognition, content-based image retrieval, face detection and recognition, and texture classification.
* Common feature extraction techniques include Histogram of Oriented Gradients (HOG), Speeded Up Robust Features (SURF), Local Binary Patterns (LBP), Haar wavelets, and color histograms.
* Reference: <http://www.sciencedirect.com/science/article/pii/S1110866513000248>

1. **Feature (in data analysis)**

* In machine learning and pattern recognition, a feature is an individual measurable property of a phenomenon being observed. Choosing informative, discriminating and independent features is a crucial step for effective algorithms in pattern recognition, classification and regression.
* Features are usually numeric, but structural features such as strings and graphs are used in syntactic pattern recognition. The concept of "feature" is related to that of explanatory variable used in statistical techniques such as linear regression.
* The initial set of raw features can be redundant and too large to be managed. Therefore, a preliminary step in many applications of machine learning and pattern recognition consists of selecting a subset of features, or constructing a new and reduced set of features to facilitate learning, and to improve generalization and interpretability.
* Extracting or selecting features is a combination of art and science; developing systems to do so is known as feature engineering. It requires the experimentation of multiple possibilities and the combination of automated techniques with the intuition and knowledge of the domain expert.
* Automating this process is feature learning, where a machine not only uses features for learning, but learns the features itself.
* Reference: <https://en.wikipedia.org/wiki/Feature_(machine_learning)>

1. **Semi-structured data**

* Semi-structured data is information that doesn’t reside in a relational database but that does have some organizational properties that make it easier to analyze. With some processing you can store them into a relational database, but the semi structure exists to ease space, clarity or compute.
* Examples of semi-structured data: CSV but XML and JSON documents are semi structured documents, NoSQL databases are considered as semi structured.
* Semi-structured data lies somewhere between the two. It is not organized in a complex manner that makes sophisticated access and analysis possible.

1. **Data Clustering**

* The idea of finding natural groupings in the data may be called unsupervised segmentation or simply clustering. Clustering is another application of our fundamental notion of similarity. The basic idea is that we want to find groups of objects (consumers, businesses, whiskeys, etc.), where the objects within groups are similar, but the objects in different groups are not so similar.
* While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.
* The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.
* Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
* Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
* In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
* Reference: <http://www.cs.cmu.edu/afs/andrew/course/15/381-f08/www/lectures/clustering.pdf>

<http://www.cs.put.poznan.pl/jstefanowski/sed/DM-7clusteringnew.pdf>

1. **Stream Mining**

* Data Stream Mining is the process of extracting knowledge structures from continuous, rapid data records. A data stream is an ordered sequence of instances that in many applications of data stream mining can be read only once or a small number of times using limited computing and storage capabilities.
* Examples of data streams include computer network traffic, phone conversations, ATM transactions, web searches, and sensor data. Data stream mining can be considered a subfield of data mining, machine learning, and knowledge discovery.
* In many data stream mining applications, the goal is to predict the class or value of new instances in the data stream given some knowledge about the class membership or values of previous instances in the data stream. Machine learning techniques can be used to learn this prediction task from labeled examples in an automated fashion.
* Data streams demonstrate several unique properties: infinite length, drift, evolution, feature-evolution and limited labeled data. Drift occurs in data streams when the underlying concept of data changes over time.
* Evolution occurs when new classes evolve in streams. Feature-evolution occurs when feature set varies with time in data streams. Data streams also suffer from scarcity of labeled data since it is not possible to manually label all the data points in the stream. Each of these properties adds a challenge to data stream mining.
* Reference: <http://link.springer.com/chapter/10.1007%2F978-3-642-29035-0_33>

1. **Granger Causality**

* Granger causality is a concept for a specific notion of causality in time-series analysis. The idea of Granger causality is a pretty simple: A variable X Granger-causes Y if Y can be better predicted using the histories of both X and Y than it can using the history of Y alone.
* Conceptually, the idea has several components:
  + Temporality: Only past values of X can cause Y.
  + Exogeneity: A necessary condition for X to be exogenous of Y is that X fails to Granger-cause Y.
  + Independence: Similarly, variables X and Y are only independent if both fail to Granger-cause the other.
* Granger causality is thus a pretty powerful tool, in that it allows us to test for things that we might otherwise assume away or otherwise take for granted.
* Reference: <http://www.scholarpedia.org/article/Granger_causality>

1. **Data Classification**

* Data classification is broadly defined as the process of organizing data by relevant categories so that it may be used and protected more efficiently. The classification process not only makes data easier to locate and retrieve – data classification is of particular importance when it comes to risk management, compliance, and data security.
* Data classification involves tagging data, which makes it easily searchable and trackable. It also eliminates multiple duplications of data, which can reduce storage and backup costs, as well as speed up the search process.
* Data classification is carried out for a variety of purposes, one of the most common being a process that supports data security initiatives. But data may be classified for a number of reasons, including ease of access, to comply with regulatory requirements, and to meet various other business or personal objectives.
* In some cases, data classification is a regulatory requirement, as data must be searchable and retrievable within specified timeframes. For the purposes of data security, data classification is a useful tactic that facilitates proper security responses based on the type of data being retrieved, transmitted, or copied.
* In the area of Data Mining, data classification is also a technique used in prediction process. It is a technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be sunny, rainy or cloudy. Popular classification techniques include decision trees and neural networks.
* Reference: <http://searchdatamanagement.techtarget.com/definition/data-classification>

<https://www.tutorialspoint.com/data_mining/dm_classification_prediction.html>

1. **Supervised Learning**

* Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

Y = f(X)

* The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.
* It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance.
* In supervised learning, each example is a pair consisting of an input object and the desired output value. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples.
* An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a reasonable way.
* Classification and Regression are some common Supervised learning methods used.

1. **Triplestore**

* A triplestore or RDF store is a purpose-built database for the storage and retrieval of triples through semantic queries. A triple is a data entity composed of subject-predicate-object.
* A key feature of many triplestores is the ability to do inference. It is important to understand that a DBMS typically offers the capacity to deal with concurrency, security, logging, recovery, and updates, in addition to loading and storing data.
* Native triplestores are those that are implemented from scratch and exploit the RDF data model to efficiently store and access the RDF data. These include: 4Store, Allegro Graph, Big Data, Jena TDB etc.
* RDBMS-backed triplestores are built by adding an RDF specific layer to an existing RDBMS. These include: Jena SDB, IBM DB2 and Virtuoso.
* NoSQL Triplestores are recently being investigated as possible storage managers for RDF. For example, CumulusRDF is built on top of Cassandra.
* Reference: http://www.dataversity.net/introduction-to-triplestores/

1. **Unsupervised learning**

* Unsupervised learning is where you only have input data (X) and no corresponding output variables.
* The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.
* These are called unsupervised learning because unlike supervised learning above there is no correct answers and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data.
* Clustering, Association, Similarity are some common unsupervised learning methods.

1. **Training Data**

* Training data is that set of data which is used to build the model for data mining processes. Generally, supervised learning datasets contain labels whereas unsupervised datasets do not contain labels.
* Based on the training set provided to the model, the model will train itself in accordance to the features of training set and then provide predictions.

1. **Test Data**

* Test data is that set of data which is used in the DM process to test the parameters of the model such as accuracy, performance etc.
* After the model is trained with training data set, the label is removed from the test set and the trained model is applied to this data.
* On comparing the predictions and actual value we come to know how well the model was trained.

1. **Clustering Analysis**

* Cluster Analysis groups data objects based on information found in the data that describes the object and their relationships. The goal is to group objects which are similar to each another and different from objects in another group. The greater the similarity within the group and the greater the difference between the groups, the better clustering.
* Clustering methods can be classified into the following categories −
  + Partitioning Method
  + Hierarchical Method
  + Density-based Method
  + Grid-Based Method
  + Model-Based Method
  + Constraint-based Method
* Reference: <https://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf>

1. **Deep learning**

* Deep Learning is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks.
* Deep learning is one of the only methods by which we can circumvent the challenges of feature extraction. This is because deep learning models are capable of learning to focus on the right features by themselves, requiring little guidance from the programmer. This makes deep learning an extremely powerful tool for modern machine learning.

1. **ETL Jobs**

* ETL is short for extract, transform, load, three database functions that are combined into one tool to pull data out of one database and place it into another database.
* Extract is the process of reading data from a database.
* Transform is the process of converting the extracted data from its previous form into the form it needs to be in so that it can be placed into another database. Transformation occurs by using rules or lookup tables or by combining the data with other data.
* Load is the process of writing the data into the target database.
* ETL is used to migrate data from one database to another, to form data marts and data warehouses and also to convert databases from one format or type to another.

1. **SQL**

* SQL stands for Structured Query Language. SQL is used to communicate with a database. According to ANSI, it is the standard language for relational database management systems.
* SQL statements are used to perform tasks such as update data on a database, or retrieve data from a database. Some common relational database management systems that use SQL are: Oracle, Sybase, Microsoft SQL Server, Access, Ingres, etc.
* Although most database systems use SQL, most of them also have their own additional proprietary extensions that are usually only used on their system. However, the standard SQL commands such as "Select", "Insert", "Update", "Delete", "Create", and "Drop" can be used to accomplish almost everything that one needs to do with a database.

1. **NoSQL**

* NoSQL database, also called Not Only SQL, is an approach to data management and database design that's useful for very large sets of distributed data.
* NoSQL, which encompasses a wide range of technologies and architectures, seeks to solve the scalability and big data performance issues that relational databases weren’t designed to address. NoSQL is especially useful when an enterprise needs to access and analyze massive amounts of unstructured data or data that's stored remotely on multiple virtual servers in the cloud.
* NoSQL DB was required because:
  + Developers are working with applications that create massive volumes of new, rapidly changing data types — structured, semi-structured, unstructured and polymorphic data.
  + Long gone is the twelve-to-eighteen-month waterfall development cycle. Now small teams work in agile sprints, iterating quickly and pushing code every week or two, some even multiple times every day.
  + Applications that once served a finite audience are now delivered as services that must be always-on, accessible from many different devices and scaled globally to millions of users.
  + Organizations are now turning to scale-out architectures using open source software, commodity servers and cloud computing instead of large monolithic servers and storage infrastructure.
* Reference: MongoDB, <http://searchdatamanagement.techtarget.com/definition/NoSQL-Not-Only-SQL>

1. **RDBMS**

* RDBMS stands for Relational Database Management System. RDBMS data is structured in database tables, fields and records. Each RDBMS table consists of database table rows. Each database table row consists of one or more database table fields.
* RDBMS store the data into collection of tables, which might be related by common fields. RDBMS also provide relational operators to manipulate the data stored into the database tables. Most RDBMS use [SQL](http://www.sql-tutorial.net/) as database query language.
* The leading RDBMS products are Oracle, IBM's DB2 and Microsoft's SQL Server.

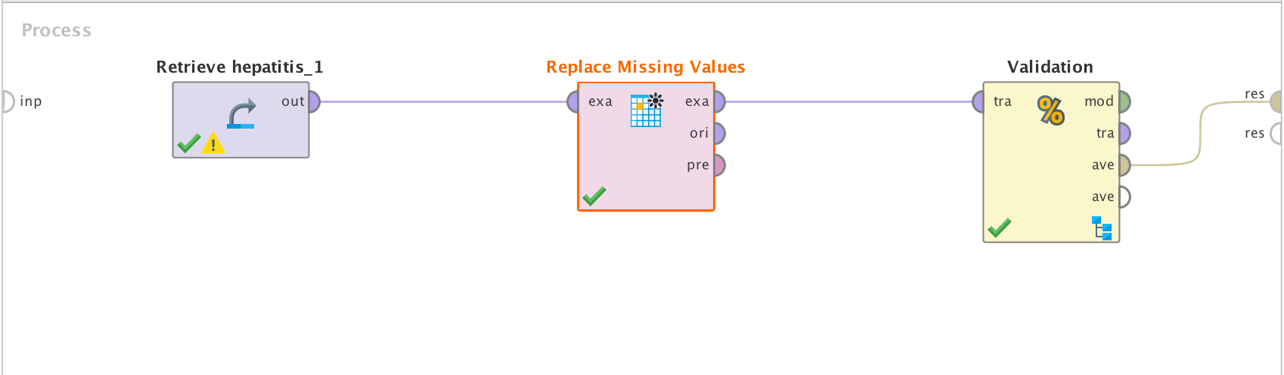
1. **Hash-joins**

* The hash join is an example of a join algorithm and is used in the implementation of a relational database management system.
* Hash join is similar to nested loop join but faster than nested loop join and hash join is used for equijoin.
* The task of a join algorithm is to find, for each distinct value of the join attribute, the set of tuples in each relation which have that value.
* Hash joins require an equijoin predicate (a predicate comparing values from one table with values from the other table using the equals operator '=').
* The classic hash join algorithm for an inner join of two relations proceeds as follows:
  + First prepare a hash table of the smaller relation. The hash table entries consist of the join attribute and its row. Because the hash table is accessed by applying a hash function to the join attribute, it will be much quicker to find a given join attribute's rows by using this table than by scanning the original relation.
  + Once the hash table is built, scan the larger relation and find the relevant rows from the smaller relation by looking in the hash table.
* The first phase is usually called the "build" phase, while the second is called the "probe" phase. Similarly, the join relation on which the hash table is built is called the "build" input, whereas the other input is called the "probe" input. It is like merge join algorithm.

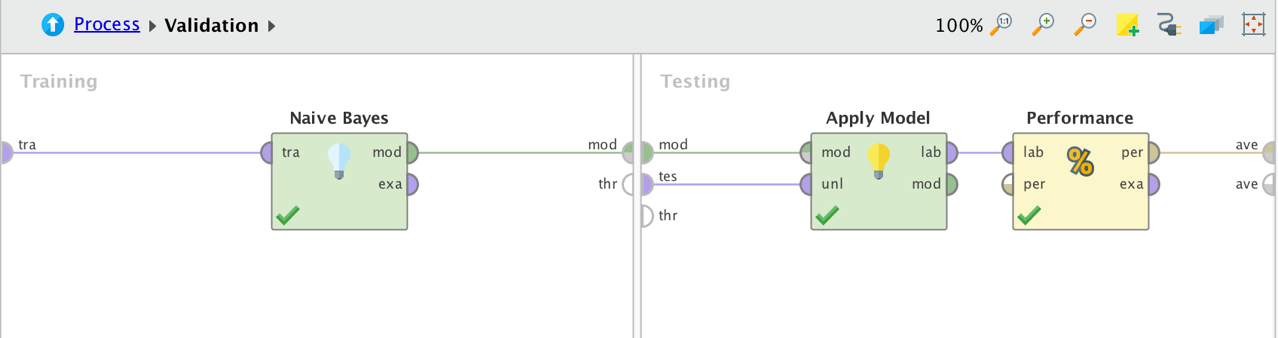
Question 2)

1.)

Below is the analytics process created in RapidMiner to work on the problem.

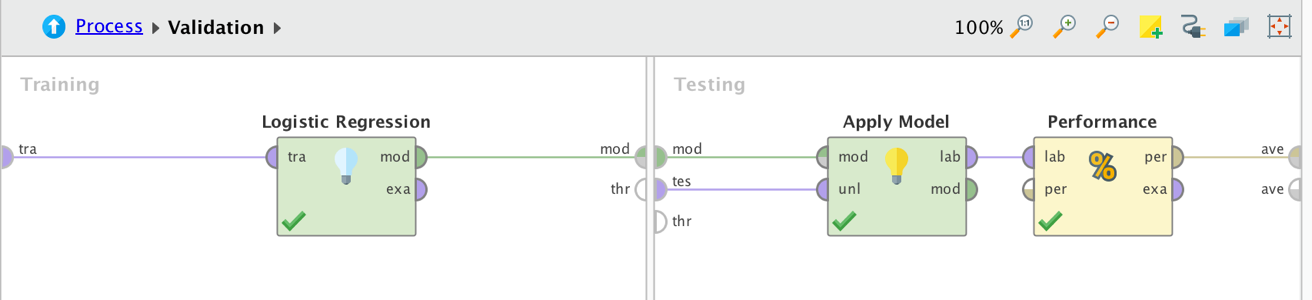


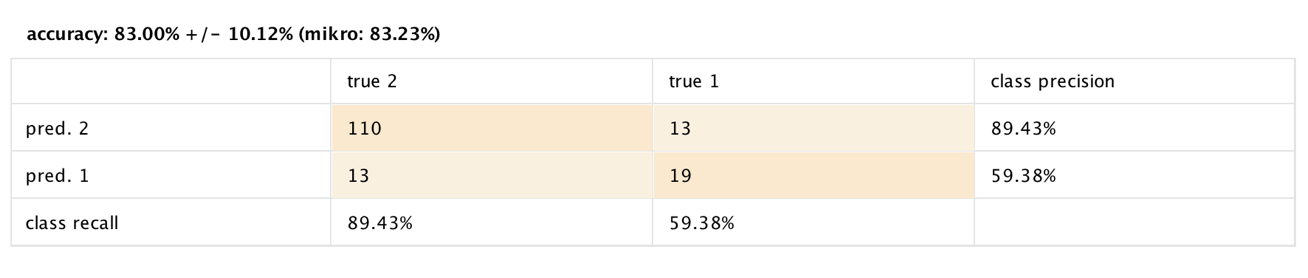
For Cross-Validation, Naïve Bayes was being used to train the model.



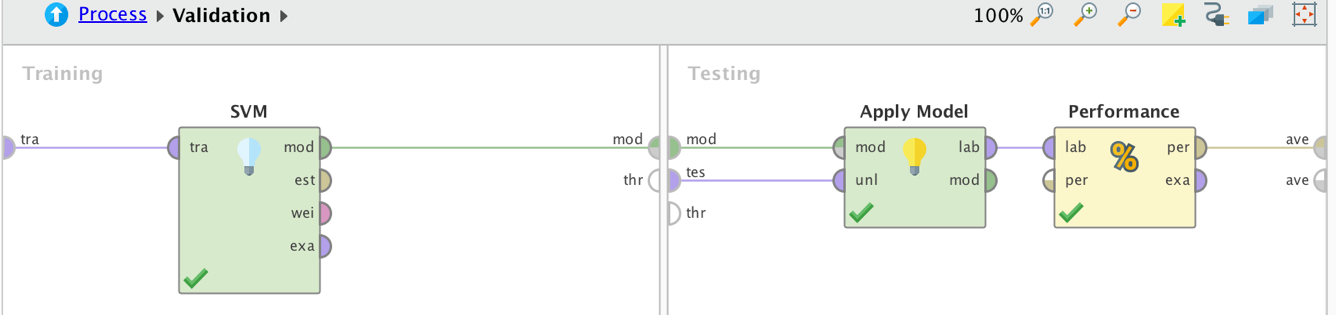
2.)

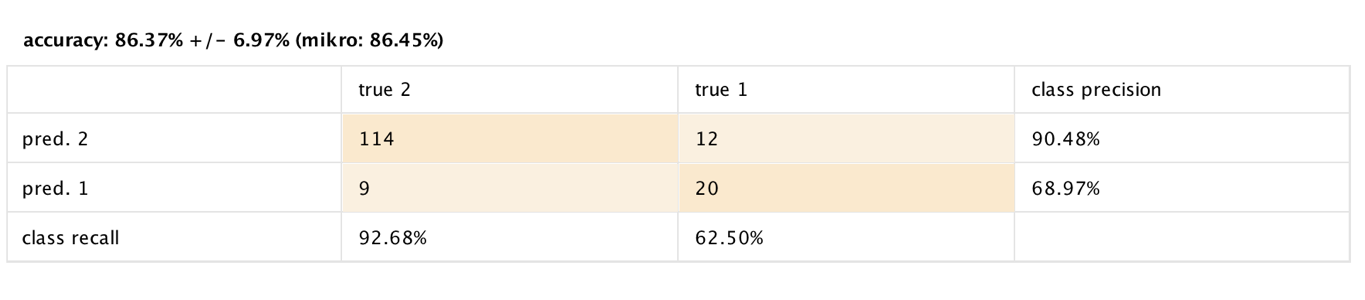
Instead of Naïve Bayes, I have used Logistic Regression for the 1st experiment. Below the accuracy and confusion matrix is given.





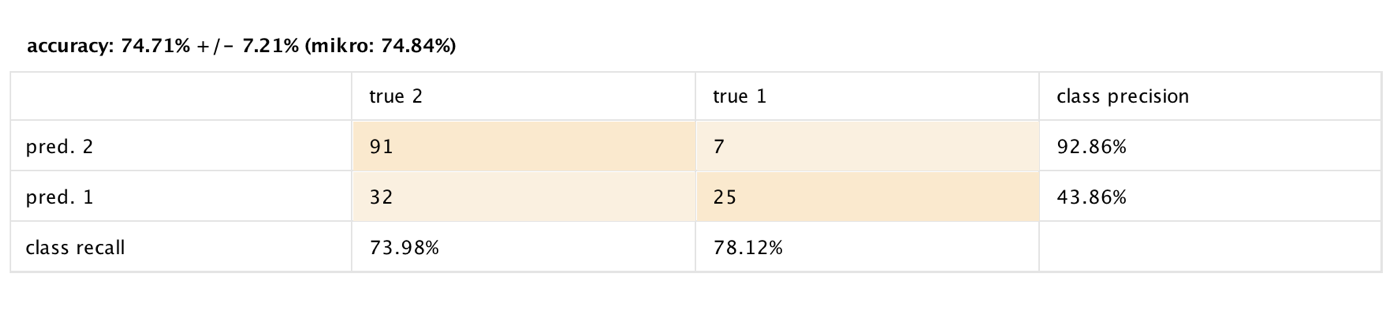
Instead of Naïve Bayes, I have used Support Vector Machines (SVM) for the 2st experiment. Below the accuracy and confusion matrix is given.





3.)

The confusion matrix created when cross-validation was done using Naïve Bayes.



4.)

In order to understand the analyze the results, we must first understand the confusion matrix and its components. Below are the details of what each component in confusion matrix denotes in our case.

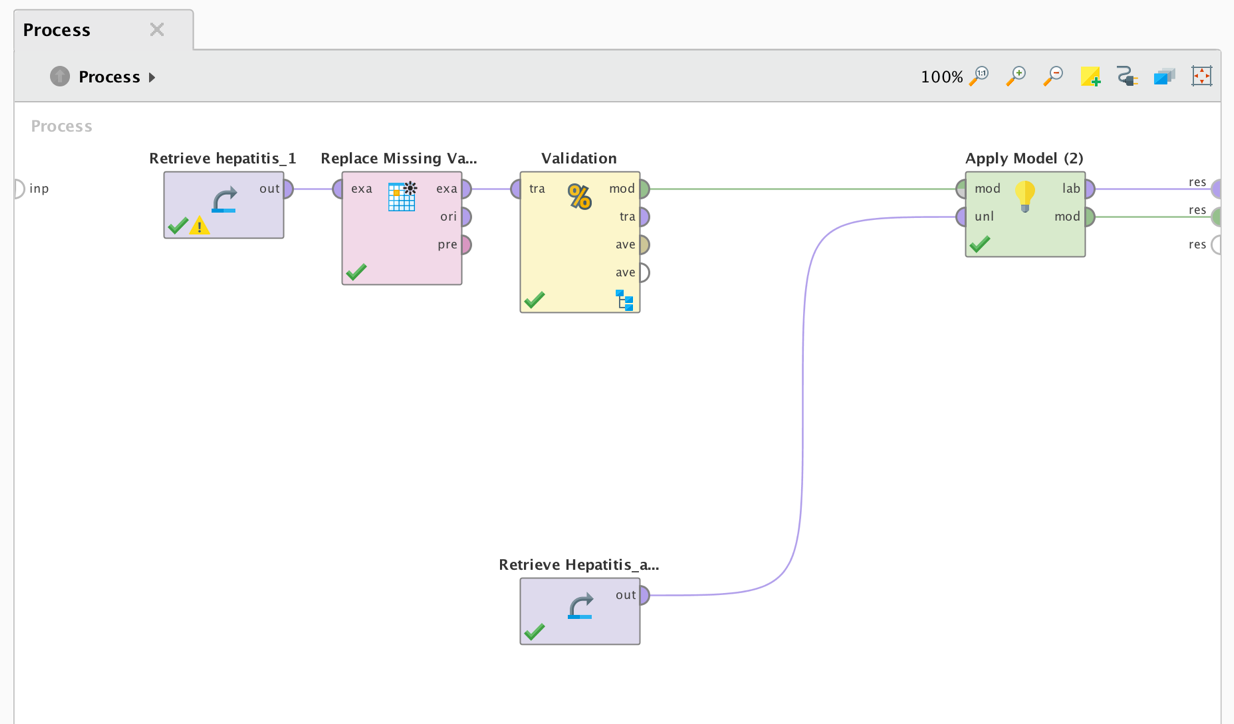
(I was not sure about the classification, so I took 1 – die and 2 – live)

* **True Positive**: Model predicted the person is going to live and that person actually lived. So the model correctly predicted that **91 persons** are going to live.
* **True Negative**: Model predicted the person is going to die and the person actually died. So the model predicted correctly that **25 persons** are going to die.
* **False Negative**: Model predicted the person is going to die but that person lived ever after. So according to the model, **32 persons** will die but they lived.
* **False Positive**: Model predicted the person is going to live but the person died. So the model predicted that **7 persons** are going to live but they died.

The accuracy of the model was about 74.71%. With that as accuracy, we can see from the matrix that there is significant number of false cases. Our aim has to be minimize the number of false cases. So, we apply different algorithms and models to get better accuracy.

This is the reason why I had included the confusion matrix of both Logistic Regression and SVM. On checking out all the confusion matrices, we can SVM performed better based on accuracy with a score of around 86%. This doesn’t mean that Naïve Bayes is not a good procedure, but for the data we have for our case, SVM works out to be better.

5.)



I created a data set with the given data set and removed labels from the file. I included that data in RapidMiner and applied it to the model as shown above.

Based on the how the model was trained, the result of the unlabeled data is shown below.

RapidMiner shows the prediction with the confidence.

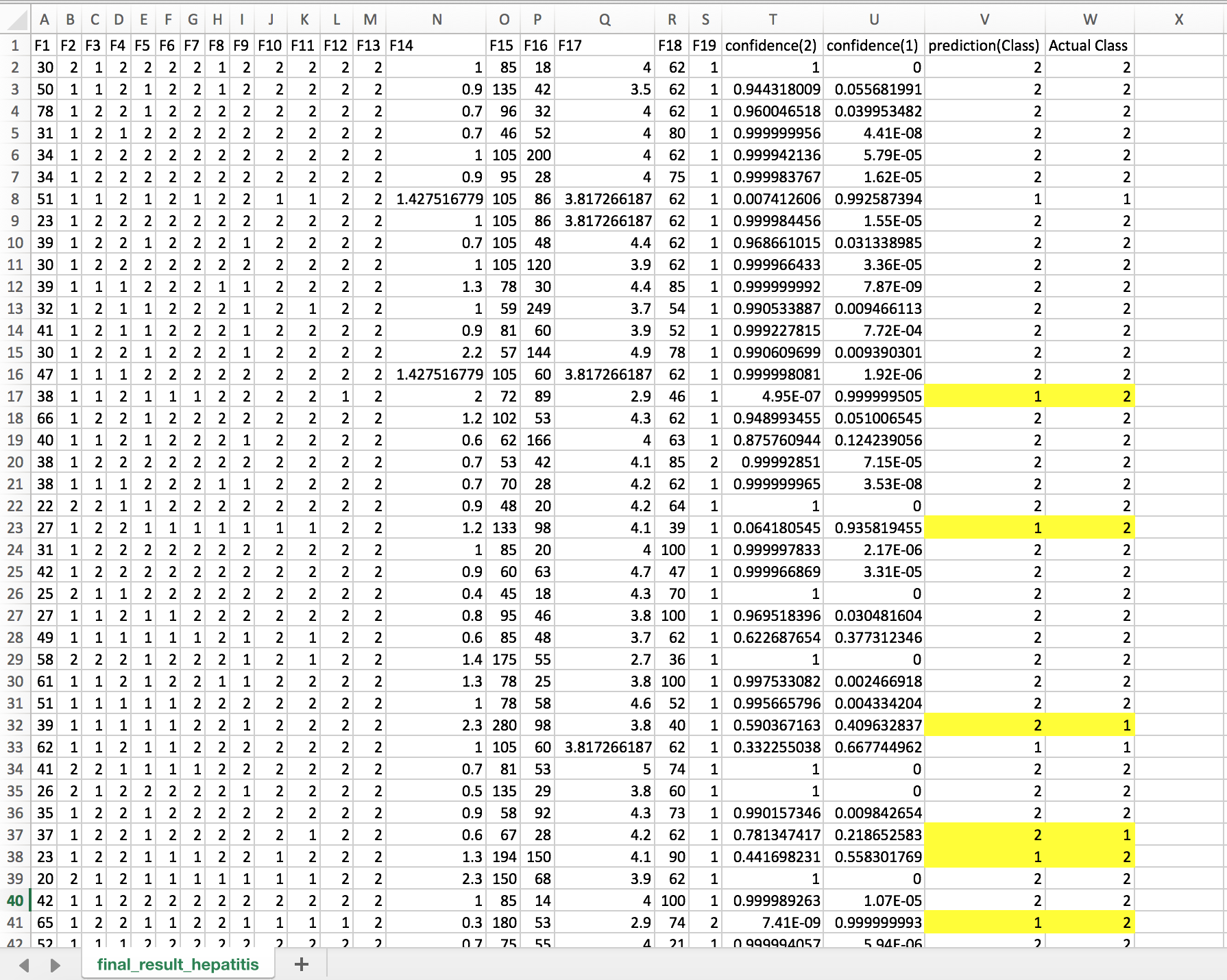


6.)

On the applying the model to the test set of 49 patients which I created, the confusion matrix came as below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  | true 2 | true 1 |  |
|  | pred. 2 | 41 | 2 |  |
|  | pred. 1 | 4 | 2 |  |
|  |  |  |  |  |

Below is the screenshot of the result which I got. The ones highlighted with yellow are the false ones.



**Question 3)**

So I researched the Target pregnancy case and I think they approached the problem in following way in CRISP-DM model.

* **Business Understanding:** It is a phase where the objective is to understand and define the business problem and convert that into a Data-Mining goal. Target Inc., to stay ahead from the rest of the market came up with a with a data-mining goal to predict whether you have a baby long before you start buying diapers. Basically they wanted to predict long before that a woman is pregnant. Target started conducting surveys which indicated that once a consumer’s shopping habits are ingrained, it can be hard to change them – except during certain brief periods of a person’s life, like after a marriage or the birth of a child, where shopping patterns and brand loyalties often change.  The birth of a child represents a new grocery and household goods list for new parents, as well as the opportunity for Target to sell things like cribs, rugs, furniture, car seats, and other items that a person or couple would not usually buy. Because birth records are public information it was already common practice for companies to send promotional items to new parents; so, to stay one step ahead of competitors, marketers at Target wanted to see if there was a way to predict pregnancy during the second trimester.
* **Data Understanding**: It involves collecting initial data, describing the data in terms of amount, type and quality of data, exploring data using available tools and verifying data quality. As pointed in the articles which I read, target started giving ‘Guest ID’ to all the guests in their store and also had their basic demographic information & e-mail address. As they were Target’s customer, they had their past transactions also with what all they bought and how did they pay. Also linked to the Guest ID is demographic information like your age, whether you are married and have kids, which part of town you live in, how long it takes you to drive to the store, your estimated salary, whether you’ve moved recently, what credit cards you carry in your wallet and what Web sites you visit. Target can buy data about your ethnicity, job history, the magazines you read, if you’ve ever declared bankruptcy or got divorced, the year you bought (or lost) your house, where you went to college, what kinds of topics you talk about online, whether you prefer certain brands of coffee, paper towels, cereal or applesauce, your political leanings, reading habits, charitable giving and the number of cars you own.
* **Data Preparation**: Data preparation (or data preprocessing) in this context means manipulation of data into a form suitable for further analysis and processing. It is a process that involves many different tasks and which cannot be fully automated. All the data gathered by Target whether they acquired it on their own or purchased it must not be of good quality. They definitely must have applied several algorithms to clean the data. Many people are not comfortable in giving all the details about themselves, so customers like these do not give all the data related to them. There may be cases where the customer can provide wrong data about themselves, so here is a challenge to identify those and remove them as data like these would lead to wrong prediction. Data preparation is essential for successful data mining. Poor quality data typically result in incorrect and unreliable data mining results. Data preparation improves the quality of data and consequently helps improve the quality of data mining results.
* **Model Planning and Building**: The modeling stage is the primary place where data mining techniques are applied to the data to generate results for the goals we want to achieve. As target had good amount of clean data to work with and build a prediction model, they started crawling the data to find out patterns in the data. Target reviewed the shopping habits of women who had a baby-shower registry as they approached their due dates. They found that Lots of people buy lotion, but they noticed that women on the baby registry were buying larger quantities of unscented lotion around the beginning of their second trimester. They also noted that sometime in the first 20 weeks, pregnant women loaded up on supplements like calcium, magnesium and zinc. Many shoppers purchase soap and cotton balls, but when someone suddenly starts buying lots of scent-free soap and extra-big bags of cotton balls, in addition to hand sanitizers and washcloths, it signals they could be getting close to their delivery date. Eventually they were able to identify about 25 different products that were indicators of pregnancy, including items like unscented lotion, vitamin supplements, hand sanitizers and washcloths. By treating the purchase of each item as a variable, they were able to create a model that assigned each shopper a pregnancy prediction score based on their purchases.
* **Evaluation**: The purpose of the evaluation stage is to assess the data mining results rigorously and to gain confidence that they are valid and reliable before moving on. If we look hard enough at any dataset we will find patterns, but they may not survive careful scrutiny. We would like to have confidence that the models and patterns extracted from the data are true regularities and not just idiosyncrasies or sample anomalies. It is possible to deploy results immediately after data mining but this is inadvisable; it is usually far easier, cheaper, quicker, and safer to test a model first in a controlled laboratory setting. Hence, the review of the model here helps us figure out the performance of model and if we find something wrong, then we can go back and again start the DM process again so that we don’t get wrong predictions.

* **Deployment**: In deployment the results of data mining—and increasingly the data mining techniques themselves—are put into real use in order to realize some return on investment. The clearest cases of deployment involve implementing a predictive model in some information system or business process. In our case, Target started applying the prediction model to the National registry and hence were able to find the thousands of woman who were pregnant and hence they would specifically target these women for baby products. Well they actually succeeded and the sales for Target’s Mom and Baby department skyrocketed.