RIGEO

REVIEW OF INTERNATIONAL GEOGRAPHICAL EDUCATION

ISSN: 2146-0353 • © RIGEO • 11(1), MARCH, 2021

www.rigeo.org

Research Article

Classify Event Participants in Universities and Industries Using Knowledge Discovery in Databases

Murnawan[1]
Widyatama University
murnawan@widyatama.ac.id

Ucu Nugraha[2]
Widyatama University

Abstract

One of the needs of event organizers is to be able to find out the profile of event participants based on the training data that has been created, find out the relationship between the selected attributes (namely, education, age, job, filed, business scale, workplace, online marketing) and the class that has been formed, and find the relationship between prospective event participants and their eligibility to attend the event. The knowledge discovery in database could be applied both in universities and industries to classify the participant's attributes. In addition, the organizers need to be able to analyze and group event participants into predetermined classes so as to facilitate the stages of the selection process. This research will analyze the data to find new knowledge using the Knowledge Discovery in Databases (KDD) stage and data mining techniques, namely Classification. Classification technique works by grouping data based on training data and the value of the classification attribute. The Classification method predicts the class of a thing from a set of attributes that describe it, where the prediction process is carried out based on a database that already has a class label first. Classifier learning is a selection process that has a high level of accuracy. Naive Bayes technique is one of the simple classification techniques but can provide a very good level of accuracy. This research will classify event participants into three classes, namely Eligible, Considered and Unfit classes based on predetermined attributes in which the participants from nearby locations with sound knowledge are preferred. One of the results of this research is expected to assist event organizers in conducting the analysis process of the relationship between event participants and the feasibility of participating in the event from the data that has been obtained. Based on the results of the performance evaluation of the classification method using the Classifier Accuracy Measurables method, it shows that the accuracy of the predictions that have been made is 94%.

Keywords

Classification, Classifier Accuracy Measurables Method, Data Mining, KDD, Naïve Bayes

To cite this article: Murnawan.; and Nugraha, U. (2021) Classify Event Participants in Universities and Industries Using Knowledge Discovery in Databases. *Review of International Geographical Education (RIGEO), 11(1), 526-542. doi: 10.48047/rigeo.11.1.36*

Submitted: 20-01-2021 • **Revised:** 15-02-2021 • **Accepted:** 25-03-2021

Introduction

With the rapid development of technology and the amount of data in the company, they must be able to process the data into useful information or the data will only become useless. This encourages companies to be able to analyze data and then extract that data into knowledge through analyzing large amounts of data, and be able to present it to be useful for users and ultimately help make the right decisions so that companies can compete with their competitors.
This research will only discuss one of the events organized by PT. XYZ is an Entrepreneurial Woman event, where this event is one of the events engaged in the service sector. They benefit from the provision of seminars and workshops with the aim of providing direction in building and developing businesses, as well as holding competitions on entrepreneurship where business actors who have registered for the competition will be monitored and assisted in developing their businesses.
The needs of the event organizer for Women Entrepreneurs are to be able to find out the profile of the contestants based on the training data that has been created, to find out the relationship between the selected attributes and the class that has been formed, and to find the relationship between the Business Actor and the eligibility to participate in the competition. In addition, the organizers need to be able to analyze and classify Business Actors into predetermined classes so as to facilitate the stages of the selection process. This research will analyze the data to find new knowledge using the Knowledge Discovery in Databases (KDD) stage and data mining techniques, namely classification. Classification technique works by grouping data based on training data and the value of the classification attribute. The grouping rules will be used to classify new data into existing groups.
The Classification method predicts the class of a thing from a set of attributes that describe it, where the prediction process is carried out based on a database that already has a class label first. Classifier learning is a selection process that has a high level of accuracy. Naive Bayes technique is one of the simple classification techniques but can provide a very good level of accuracy. This research will classify Business Actors into three classes, the Business Actor is in the Eligible category in the sense that it is potentially eligible to take part in the competition (according to the competition criteria), considered (could be appropriate, could not be) or Unfit (not in accordance with the competition criteria). The results of this research are expected to be able to find new knowledge (Knowledge Discovery) in the form of a profile of the classification results where this can assist in evaluating and supporting PT.XYZ's business processes. In addition, it can assist the organizers of the Women Entrepreneurs Event in conducting the analysis process of the relationship between Business Actors and the feasibility of participating in the competition from the data that has been obtained.

Literature Review

Knowledge Discovery in Database

Knowledge discovery in database (KDD) is a model for obtaining knowledge from an existing database, this knowledge can be in the form of a decision-making need (Kapoor, 2010). According to Max Bram, Knowledge discovery in database (KDD) is defined as extraction of data to find previously unknown and potentially useful information (Işık, Jones, & Sidorova, 2013). Knowledge discovery in databases has a step-by-step process that must be carried out (Laudon, 2007).

Selection

Data from a set of operational data needs to be done before the stage data mining in KDD begins. Data from the selection results that will be used for the data mining process, is stored in a file, separate from the operational database.

Pre-processing or Cleaning

Before the data stage mining can be carried out, it is necessary to perform a stage of the pre-processing or cleaning data. The stage pre-processing or cleaning includes removing duplicate data, checking for inconsistent data, and correcting errors in data, such as typographical errors.

In the stages preprocessing and cleaning and process can be carried enrichment out, namely the process of "enriching" existing data with data or other information that is relevant and needed for KDD, such as external data or information. The preprocessing or cleaning stage is a very important stage, because the stage preprocessing or cleaning has a percentage of 80% of the rules that will be generated in data mining (Laudon, 2007).

Transformation

stage transformation is a stage in grouping data into categorical data. This data grouping is done so that the data is suitable to be carried out in the stage data mining. The stage is transformation very dependent on the type or pattern of information to be searched for in the database.

Data Mining

Data mining is the stage of looking for interesting patterns or information in selected data using certain techniques or methods. Techniques, methods, or algorithms in data mining vary widely. Selection of the right method or algorithm is very dependent on the objectives and process of KDD as a whole.

Interpretation or Evaluation

The evaluation stage is the stage to see the accuracy value of the patterns that have been generated in the process data mining. In addition, information patterns generated from the data process mining also need to be displayed in a form that is easily understood by interested parties. This stage is part of the KDD process called interpretation.

Data Mining

Definition of Data Mining

Data mining or better known as Knowledge Discovery in Database (KDD) is a field of several scientific fields that unites patterns, statistics, databases and visualization to solve problems from databases large (Kimball & Ross, 2010). According to the Gartner Group, data mining is the process of finding correlations, patterns and trends meaningful new by sorting large amounts of data in data storage using pattern recognition technology as well as statistical and mathematical techniques. Data mining is an analysis that is carried out automatically on large and complex data with the aim of obtaining important patterns whose existence is not realized (Vercellis, 2011).

Methods Data Mining

In general methods are data mining divided into 6 methods including (Vercellis, 2011):

Description

Description method is a method that aims to find a way to describe the patterns and trends that exist in a data. Pattern descriptions and trends can provide a possible explanation for a pattern or trend (Marakas & O'Brien, 2013).

Classification

The classification method is a method that classifies data that has or sets input or predictor attributes and target attributes that are categorical. Meaning that is categorical is like an attribute of income groups which can be separated into low income, medium income and high income. Algorithms that can be used in the classification method are k-nearest neighbor, decision tree, naïve Bayes algorithm and algorithm neural network (Robbins & Coulter, 2012).

Estimation

The estimation method has the same method as the classification method; however, the estimation method has a numeric target attribute while the classification method has a categorical attribute. The estimation model is constructed using a complete record that provides the value of the target attribute as a predictive value. Furthermore, in the next review, the estimated value of the target attribute is made based on the predicted attribute value (Turban, Sharda, Delen, & King, 2011).

Prediction

The prediction method also has the same method as the method of classification and estimation. However, the results of the predicted value will be visible in the future. There are several algorithms that use classification and estimation methods that can be used in prediction methods but in suitable circumstances. These algorithms include simple linear regression and correlation algorithms, multiple regression, neural network, decision tree and k-nearest neighbor (Person, 2013).

Clustering

The clustering method is a grouping of records, observations or grouping cases into classes that have similar cases. Clustering is a different method from other classification methods because the clustering method only has a set of input or predictor attributes and does not have a target attribute. This is because the clustering algorithm serves to segment the entire data set into subgroups or clusters based on the similarities that the data has.

Association

The association method is a method that aims to generate rules for measuring the relationship between two or more attributes. Algorithms that can be used to produce an association rule are the a priori algorithm and the GRI algorithm.

Classification Methods

Classification methods are included in the method supervised learning. This is because the classification method uses a set of data with a known label or class so that the data contains predictor attributes and target attributes. The classification method is divided into 2 stages (Vercellis, 2011):

The Stage of Making the Classification Model

At the stage of making the classification model the data used is data training. Data Training is data that has been specified label or classification class data is Training used as sample data in making a classification model. The classification model creation stage is also referred to as the induction process.

The Stage of Applying the Classification Model.

The data used in this stage are data testing. In the data classification method testing used is data supervised learning. Data that is supervised learning is data that already has a label or class classification is known. The application stage of the classification model aims to predict the class classification of the data in the data testing based on the classification model that has been made. The stage of applying the classification model is also known as the deduction process (Cegielski & Rainer, 2009).

Naïve Bayes

The naïve bayes algorithm is a simple probabilistic classification algorithm that performs calculations by adding up the frequencies and value combinations from dataset a given [16].

theory Bayes was first discovered by Thomas Bayes in 1950 which aims to predict future events with the proviso previous events that have occurred (Connolly & Begg, 2005). In general, the Bayes theory is written in an equation (1) (Inmon, 2005).

$$P(H|X) = \frac{P(X|H)P(H|)}{P(X)} \qquad (1)$$

Description:

X : Data with unknown class
H : Hypothesis of data which is a specific class
P (H | X) : Probability of hypothesis H based on condition X (posteriori probability)
P (H) : Hypothesis probability H (Prior probability)
P (X | H) : Probability X based on the conditions in the hypothesis H
P (X) : Probability X

In determining the suitable class for the analyzed sample, a description of the theory is carried out naïve bayes. Therefore, the theory naïve Bayes adjusted can produce equation (2).

$$P(C|F_1 \dots F_n) = \frac{P(C)P(F_1 \dots F_n|C)}{P(F_1 \dots F_n)} \qquad (2)$$

Description:

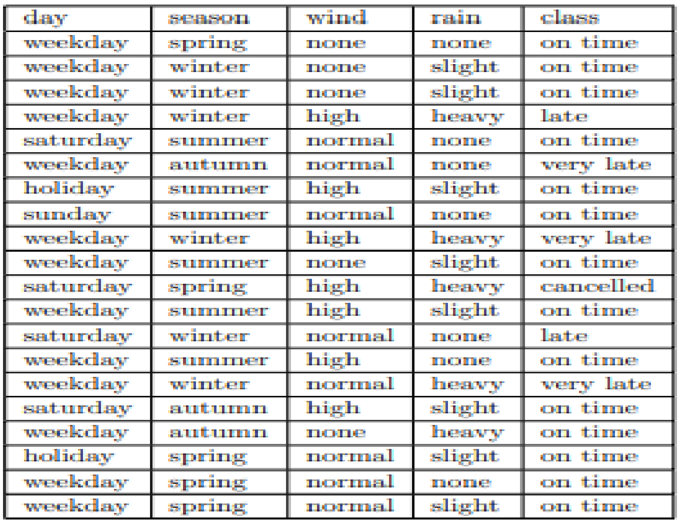
C : Class or attribute
$F_1 \dots F_n$: Class characteristic or attribute
$P(C|F_1 \dots F_n)$: A sample of certain characteristics in class C (Posterior)
$P(C)$: Chance of emergence of class C (Prior)
$(F_1 \dots F_n|C)$: Chance of appearing of sample characteristics in class C (likelihood)
$P(F_1 \dots F_n)$: Opportunity for the emergence of sample characteristics globally (evidence)

Based on the information above, the equation for the theory naïve bayes can be written simply by equation (3).

$$Posterior = \frac{Prior \text{ x } Likelihood}{Evidence} \qquad (3)$$

The requirement for a probability is that each probability has a value of 0 to 1 and the sum of all probabilities must have a total value of 1. In the classification process, the dataset used is like the data training below:

TABLE I
Sample Data Training



day	season	wind	rain	class
weekday	spring	none	none	on time
weekday	winter	none	slight	on time
weekday	winter	none	slight	on time
weekday	winter	high	heavy	late
saturday	summer	normal	none	on time
weekday	autumn	normal	none	very late
holiday	summer	high	slight	on time
sunday	summer	normal	none	on time
weekday	winter	high	heavy	very late
weekday	summer	none	slight	on time
saturday	spring	high	heavy	cancelled
weekday	summer	high	slight	on time
saturday	winter	normal	none	late
weekday	summer	high	none	on time
weekday	winter	normal	heavy	very late
saturday	autumn	high	slight	on time
weekday	autumn	none	heavy	on time
holiday	spring	normal	slight	on time
weekday	spring	normal	none	on time
weekday	spring	normal	slight	on time

Classifier Accuracy Measurables

Classifier Accuracy Measures (Han and Kimber, 2006: 360) is a classification method based on the accuracy of the model in making predictions. This is done because accuracy in processing data is one of the important things (Pardillo & Mazón, 2011). The method used to test the accuracy of this classification model is the method hold out. In this method, the original data is partitioned into two separates called the sets training set and the test set. The classification model is then built on the basis of the training set and the results are then evaluated using the testing set. The accuracy of each classification method can be estimated based on the accuracy obtained from the test set. The proportion between the training set and the test set is not binding, but so that the variance in the model is not too large, it can be determined that the proportion of the training set is greater than the test set. Usually, 2/3 of the data is used as a training set and 1/3 is used as a testing set. The size of the accuracy of a classifier can be determined using calculations Classifier Accuracy Measurables as follows:

$$Sensivity = \frac{tpos}{pos} \qquad (4)$$

$$Specificity = \frac{tneg}{neg} \qquad (5)$$

$$Precision = \frac{tpos}{(t_{pos}+f_{pos}} \qquad (6)$$

$$Accuracy = sensivity \frac{pos}{(pos+neg)} + specifity \frac{neg}{(pos+neg)} \qquad (7)$$

t_pos is the number of true positives, namely the number of data successfully predicted by the classifier correctly (for example the number of data class "yes" from the sample can be predicted correctly by the classification model), pos. is the number samples of data positives ("yes"), t_neg is the number of true negatives that are the opposite of true positives (i.e. the amount of data class "no" from the samples properly can be predicted properly by the classification model), neg is the total number of samples negatives ("No"), and f_pos are false positives, namely the number of incorrect data predicted by the classifier ("no" is predicted as "yes").

Sensitivity is a measure of the degree to which the classifier can recognize positive samples ("yes") based on the number of true positives that can be predicted correctly if the sample are given positives. Specificity is a measure of the degree to which the classifier can recognize negatives samples ("no") based on true negatives which can be predicted correctly if given sample negatives. Precision is the percentage of the classifier in correctly guessing the true positives ("yes") class by looking at the ratio true positive that can be predicted by the sum of true positives and false positives. Accuracy is the degree of measurement which is a function of the Sensitivity and Specificity of the classification model in making predictions.

Discussion and Result

Training Data

Training Data obtained from data sources that determine the feasibility of which is considered the best by the company, the feasibility is still determined only by individual intuition. Where the training data contains 112 records, with varying attribute criteria. There are nine attributes which are divided into two criteria, namely main attributes and supporting attributes.

TABLE II
Main Attributes

Attribute	Description
Age Range	The age range of business actors, namely, from 20 years old to 50 years old.
Business Type	Type of business is being run, consisting of 3 (three) categories, Main Business, Side or Seasonal.
Business Length	A description of how long business actors have run their business, divided into three time periods, namely <1 year, 1-2 years and >2 years
Business Scale	An explanation of how big the business scale of the business-by-business actors, ranging from <200 million to >50billion

TABLE III
Supporting Attributes

Attribute	Description
Educational	Educational background of business actors, consisting of SMA, D3, S1, to S2
Main Job	Description of the work being carried out by business actors, such as housewives, employees, entrepreneurs, or students.
Business Field	Description of what type/field of business is being carried out by business actors, such as services, trade, producers and suppliers
Workforce	Description of how many workers is employed by business actors, ranging from 4 to > 100 people
Online Marketing	Description of whether the business actor has marketed their products through online media or not

Identification Of Attributes to Be Analyzed

The attributes and relationships between attributes that will be analyzed in the process mining include:

1. The relationship between the eligibility of the participant and age range
2. The relationship between the eligibility of the participant and education
3. The relationship between the eligibility of the participant and main job
4. The relationship between the eligibility of the participant and business type
5. The relationship between the eligibility of the participant and business field
6. The relationship between the eligibility of the participant and business length
7. The relationship between the eligibility of the participant and business scale
8. The relationship between the eligibility of the participant and workforce.
9. The relationship between the eligibility of the participant and online marketing

Cleaning & Integration

The first step in the Knowledge Discovery in Databases (KDD) stage is through the data stage cleaning. The process carried out is by removing noise and "null" data, removing duplicate data and correcting errors in data such as writing errors. The stage integration has been carried out on the data sources that have been obtained, but this tool can also integrate newly entered data with existing data. The purpose of data cleaning is that the data to be processed in data mining is clean and quality data. At this writing the data that will be used in the cleaning process is sourced from the Business Actor Table. The data is still not clean so the data cleaning process is carried out.

Data Selection

Based on the data training and after carrying out the stage Cleaning and Integration, at this stage the attributes that will be used in the data mining process include:

1. The city attribute is used for the selection of the data process mining based on the City of origin of the Business Actor.
2. The Age Range attribute is used for the process mining to determine the relationship between eligibility to participate in the competition and the age of the business actor concerned.
3. The education attribute is used for the process mining last in order to determine the relationship between eligibility to participate in the competition and the latest education that has been undertaken by the business actor concerned.
4. The main job attribute is used for the process mining to determine the relationship between eligibility to participate in the competition and the main job being undertaken by the business actor involved. concerned.
5. The business type attribute is used for the process mining to determine the relationship between eligibility to participate in the competition and the type of business being undertaken by the business actor concerned, whether the business is a main, side or seasonal business.
6. The business length attribute is used for the process mining to determine the relationship between the eligibility to participate in the competition and the length of business that the business actor has undertaken.
7. The business field attribute is used for the process mining to determine the relationship between eligibility to participate in the competition and the business field being undertaken by the business actor concerned, whether the business is in the field of Services, Trade, Producers or Suppliers.
8. The business scale attribute is used for the process mining to determine the relationship between the eligibility to participate in the competition and the business scale of the business that has been undertaken by the business actor concerned.
9. The workforce attribute is used for the process mining to determine the relationship between eligibility to participate in the competition and the number of workers owned by the business actor concerned.
10. The online marketing attribute is used for the process mining to determine the relationship between eligibility to participate in the competition and whether business actors have done online marketing or not.

Data Transformation

Data transformation is the process of transforming data into the form required for the process data mining next. So, at this stage, some changes are made to the attributes whose shape is still not suitable for excavation.

At first the data source is still in numeric form, then will be changed according to the form based on the attribute values and data types that are suitable for the process mining.

Data Mining

The data mining stage is the stage where the process of looking for patterns or interesting information is used. The technique used is naive bayes classification by calculating the probability of class membership.

The following is an explanation of the steps of the naive bayes classification technique: The table used is the result (training data) table where the table has a total of 112 records and previously had predetermined classes. Classes that will be formed are eligible, considered and unfit

The following is the estimated value from the training data currently available:

a. Probability of Eligible class from existing data:

P(Eligible) = 36/112

b. Probability of class Considered from existing data:

P(Considered)= 40/112

c. Probability of class Unfit from existing data:

P(Unfit) = 36/112

Probability of the Age Range attribute to the Eligible class:

TABLE IV
Eligible Class for Age Range Attributes

Age Range	P (20 – 24 | Eligible) = 0/36
	P (25 – 29 | Eligible) = 2/36
	P (30 – 34 | Eligible) = 14/36
	P (35 – 39 | Eligible) = 10/36
	P (40 – 44 | Eligible) = 10/36
	P (45 – 49 | Eligible) = 0/36
	P (>50 | eligible) = 0/36

Probability of the attribute Age Range against the class Considered:

TABLE V
Considered Class for Age Range Attributes

Age Range	P (20 – 24 | considered) = 0/40
	P (25 – 29 | considered) = 4/40
	P (30 – 34 | considered) = 19/40
	P (35 – 39 | considered) = 9/40
	P (40 – 44 | considered) = 8/40
	P (45 – 49 | considered) = 0/40
	P (>50 | considered) = 0/40

Probability of the attribute Age Range against the Unfit class:

TABLE VI
Unfit Class for Age Range Attributes

Age Range	P (20 – 24 | Unfit) = 13/36
	P (25 – 29 | Unfit) = 1/36
	P (30 – 34 | Unfit) = 6/36
	P (35 – 39 | Unfit) = 8/36
	P (40 – 44 | Unfit) = 2/36
	P (45 – 49 | Unfit) = 3/36
	P (>50 | Unfit) = 3/36

Probability Education attribute to Eligible class:

TABLE VII
Eligible Class for Education Attributes

Education	P (SMA | Eligible) = 2/36
	P (D3 | Eligible) = 7/36
	P (S1 | Eligible) = 23/36
	P (S2 | Eligible) = 4/36
	P (S3 | worth) = 0/36

Probability Education attribute to Considered class:

TABLE VIII
Considered Class for Education Attributes

Education	P (SMA | considered) = 5/40 P (D3 | considered) = 4/40 P (S1 | considered) = 27/40 P (S2 | considered) = 4/40 P (S3 | considered) = 0/40

Probability Education attribute to Unfit class:

TABLE IX
Unfit Class for Education Attributes

Education	P (SMA | Unfit) = 10/36 P (D3 | Unfit) = 4/36 P (S1 | Unfit) = 18/36 P (S2 | Unfit) = 4/36) P (S3 | Unfit) = 0/36

Probability Main Job attribute to Eligible class:

TABLE X
Eligible Class for Main Job Attributes

Main Job	P (Employee | Eligible) = 6/36 P (entrepreneur | Eligible) = 27/36 P (housewife | Eligible) = 3/36 P (student | Eligible) = 0 /36

Probability Main Job attribute to Considered class:

TABLE X
Considered Class for Main Job Attributes

Main Job	P (Employee | considered) = 14/40 P (entrepreneur | considered) = 23/40 P (housewife | considered) = 3/40 P (student | considered) = 0/40

Probability Main Job attribute to Unfit class:

TABLE XI
Unfit Class for Main Job Attributes

Main Job	P (Employees | unfit) = 15/36 P (entrepreneurs | unfit) = 14/36 P (housewife | unfit) = 2/36 P (student | Unfit) = 5/36

Probability Business Type attribute to Eligible class:

TABLE XII
Eligible Class for Business Type Attributes

Business Type	P (main | Eligible) = 23/36
	P (side | Eligible) = 13/36
	P (seasonal | Eligible) = 0/36

Probability Business Type attribute to Considered class:

TABLE XIII
Considered Class for Business Type Attributes

Business Type	P (main | considered) = 17/40
	P (sideline | considered) = 20/40
	P (seasonal | considered) = 3/40

Probability Business Type attribute to Unfit class:

TABLE XIV
Unfit Class for Business Type Attributes

Business Type	P (main | Unfit) = 16/36
	P (side | Unfit) = 17/36
	P (seasonal | Unfit) = 3/36

Probability Business Length attribute to Eligible class:

TABLE XV
Eligible Class for Business Length Attributes

Business Length	P (< 1 year | feasible) = 0/36
	P (1 1-2 years | feasible) = 3/36
	P (>2 years | feasible) = 33/36

Probability Business Length attribute to Considered class:

TABLE XVI
Considered Class for Business Length Attributes

Business Length	P (< 1 year | considered) = 9/40
	P (1-2 years | considered) = 21/40
	P (>2| considered) = 10/40

Probability Business Length attribute to Unfit class:

TABLE XVII
Unfit Class for Business Length Attributes

Business Length	P (< 1 year | Unfit) = 16/36
	P (1-2 years | Unfit) = 8/36
	P (2 years >| Unfit) = 12/36

Probability Business Field attribute to Eligible class:

TABLE XVIII
Eligible Class for Business Field Attributes

Business Field	P (Service | Eligible) = 21/36 P (Producer | Eligible) = 8/36 P (Supplier | Eligible) = 0/36 P (Trade | Eligible) = 7/36

Probability Business Field attribute to Considered class:

TABLE XIX
Considered Class for Business Field Attributes

Business Field	P (Services | Considered) = 13/40 P (Producer | Considered) = 13/40 P (Supplier | Considered) = 4/40 P (Trade | Considered) = 10/40

Probability Business Field attribute to Unfit class:

TABLE XX
Unfit Class for Business Field Attributes

Business Field	P (Service | Unfit) = 13/36 P (Producer | Unfit) = 14/36 P (Supplier | Unfit) = 2/36 P (Trade | Unfit) = 7/36

Probability Business Scale attribute to Eligible class:

TABLE XX1
Eligible Class for Business Scale Attributes

Business_Scale	P (Micro | Eligible) = 25/36 P (Small | Eligible) = 11/36 P (Medium 1 | Eligible) = 0/36 P (Medium 2 | Eligible) = 0/36 P (Large | Eligible) = 0/36

Probability Business Scale attribute to Considered class:

TABLE XXII
Considered Class for Business Scale Attributes

Business_Scale	P (Micro | Considered) = 29/40 P (Small | Considered) = 9/40 P (Medium 1 | Considered) = 2 /40 P (Medium 2 | Considered) = 0/40 P (Large | Considered) = 0/40

Probability Business Scale attribute to Unfit class:

TABLE XXIII
Unfit Class for Business Scale Attributes

Business Scale	P (Micro | Unfit) = 32/36
	P (Small | Unfit) = 3/36
	P (Medium 1 | Unfit) = 0/36
	P (Medium 2 | Unfit)) = 0/36
	P (Large | Unfit) = 1/36

Probability Workforce attribute to Eligible class:

TABLE XXIV
Eligible Class for Workforce Attributes

Workforce	P (1 - 4 | Eligible) = 10/36
	P (5-19 people | Eligible) = 17 /36
	F (20 to 100 people | Eligible) = 8/36
	F (>100 people | Eligible) = 1/36

Probability Workforce attribute to Consider class:

TABLE XXIV
Consider Class for Workforce Attributes

Workforce	P (1 - 4 people | Consider) = 29/40
	P (5-19 people | Consider) = 10/40
	P (20 s / d 100 | Consider) = 1/40
	P (> 100 | consider) = 0/40

Probability Workforce attribute to Unfit class:

TABLE XXV
Unfit Class for Workforce Attributes

Workforce	P (1 - 4 | Unfit) = 29/36
	P (5-19 people | Unfit) = 7/36
	P (20 to 100 people | Unfit) = 0/36
	F (>100 people | Unfit) = 0/36

Probability Online Marketing attribute to Eligible class:

TABLE XXVI
Eligible Class for Online Marketing Attributes

Online Marketing	P (Yes | Eligible) = 23/36
	P (No | Eligible) = 13/36

Probability Online Marketing attribute to Consider class:

TABLE XXVII
Consider Class for Online Marketing Attributes

Online Marketing	P (Yes | Consider) = 31/40 P (No | consider) = 9/40

Probability Online Marketing attribute to Unfit class:

TABLE XXVIII
Unfit Class for Online Marketing Attributes

Online Marketing	P (Yes | Unfit) = 20/36 P (No | Unfit) = 16/36

Pattern Evaluation

After carrying out the data mining process, the next stage is pattern evaluation. In this process an evaluation of the results of the mining process that has been carried out is carried out.
In the tools used, the data source is taken from the business actor entity. The purpose of data collection is to find out the profiles of potential contestants and what classes are suitable for them. The attributes used in the selection are based on the relationship with the analysis carried out. The attributes used are Age Range, business type, business length, business scale, education, occupation, business field, workforce number, online marketing. This data mining process uses the Nave Bayes Classification technique. Where the output of the process is three classes, namely 'Eligible', 'Considered', 'Unfit'. The following are the results of the evaluation of the data mining process in the form of typical patterns from the potential participants in the competition, namely in the form of profiles where this is related to attribute analysis carried out based on the classes that have been formed.

Knowledge Presentation

After all stages of the process have been mining carried out, the last stage in KDD is the Knowledge presentation stage. This stage aims to present the objectives and visualize the results that have been produced in the process mining. This knowledge presentation will be used to present the information that has been generated in the process mining before it will be used by the user.
All of the participants in the competition, the age range of 30 – 34 years is more dominant in the Eligible and Considered class, while the Unfit class is in the 35 – 39-year-old class.

a. From all participants in the competition, S1 education was more dominant in all classes.
b. Of all the participants in the competition, the number of main jobs is more dominant in the Considered and Unfit class, while the Eligible class is in the Entrepreneurship field.
c. Of all the participants in the competition, the business type is more dominant in the Eligible and Unfit class.
d. Of all the participants in the competition, business length more than 2 Years is more dominant in the Eligible class, while the Considered class is 1 - 2 year and Unfit at 1 year.
e. Inappropriate From the overall competition participants, business field are more dominant in the Eligible and Unfit classes, while Business Types are 2 Producers in the Considered class.
f. From the whole race, business scale micro more dominant numbers in all classes.
g. Of all the participants in the competition, the total number of workers is 5 to 19 The number of workers is more dominant in the Eligible class, while the Considered and Unfit class is 4 peoples
h. Out of all participants in the competition, Business Actors who do not conduct online Marketing (No) are more dominant in the Eligible and Unfit classes, while the Considered Class is for Business Actors who do online Marketing (Yes).

Calculation of Classifier Accuracy Measurables

Number of data sources before Cleaning: 1,570 Records
Number of data sources after Cleaning: 1,233 Records
Where the data sources will be divided with the provisions of 2/3 of the data sources being training data, and 1/3 of the data sources being testing data.
Training Data: 1,233 x (2/3) = 822 Record
Testing Data: 1,233 x (1/3) = 411 Record

Confusion Matrix

This confusion matrix is created by forming three matrices from predetermined classes, namely the Eligible matrix, the Considered matrix and Unfit matrix. Where the matrices are as follows:

TABLE XXIX
Eligible Confusion Matrix

		Predict	
		Eligible	**! Eligible**
Actual	Eligible	96	0
	! Eligible	6	309

T_pos (Eligible) = 96 Record
F_pos (Eligible) = 6 Record
T_neg (Eligible) = 309 Record
F_neg (Eligible) = 0 Record
Total Data Testing = 411 Record

Senstivity (Eligible) $= \frac{\text{t pos}}{\text{pos}} = \frac{96}{96} = 1$

Specificity (Eligible) $= \frac{\text{t neg}}{\text{neg}} = \frac{309}{315} = 0.98$

Precision (Eligible) $= \frac{\text{t\_pos}}{(\text{t\_pos + f\_pos})} = \frac{96}{(96+6)} = 0.94$

Accuracy (Eligible) = sensitivity $\frac{\text{pos}}{(\text{pos + neg})}$ + specifity $\frac{neg}{(pos + neg)}$

$= 1 \frac{96}{(96+315)} + 0.98 \frac{315}{(96+315)}$ =0.97 x 100% = 97 %

TABLE XXX
Considered Confusion Matrix

		Predict	
		Considered	**! Considered**
Actual	Considered	198	2
	! Considered	32	179

T_pos (Considered) = 198 Record
F_pos (Considered) = 32 Record
T_neg (Considered) = 179 Record
F_neg (Considered) = 2 Record
Total Data Testing = 411 Record

Senstivity (Considered) $= \frac{\text{t\_pos}}{\text{pos}} = \frac{198}{200} = 0.99$

Specificity (Considered) $= \frac{t\_neg}{neg} = \frac{179}{211} = 0.84$

Precision (Considered) $= \frac{t\_pos}{(t\_pos + f\_pos)} = \frac{198}{(198+32)} = 0.86$

Accuracy (Considered) = sensitivity $\frac{pos}{(pos + neg)}$ + specifity $\frac{neg}{(pos + neg)}$

$= 0.99 \frac{200}{(200+211)} + 0.84 \frac{211}{(200+211)} = 0.89$ x 100% = 89 %

TABLE XXX
Unfit Confusion Matrix

		Predict **Unfit**	 **! Unfit**
Actual	Unfit	79	36
	! Unfit	0	296

T_pos (Unfit) = 79 Record
F_pos (Unfit) = 0 Record
T_neg (Unfit) = 296 Record
F_neg (Unfit) = 36 Record
Total Data Testing = 411 Record

Senstivity (Unfit) $= \frac{t\_pos}{pos} = \frac{79}{115} = 0.68$

Specificity (Unfit) $= \frac{t\_neg}{neg} = \frac{296}{296} = 1$

Precision (Unfit) $= \frac{t\_pos}{(t\_pos + f\_pos)} = \frac{79}{(79+0)} = 1$

Accuracy (Unfit) = sensitivity $\frac{pos}{(pos + neg)}$ + specifity $\frac{neg}{(pos + neg)}$

$= 1 \frac{115}{(115+296)} + 0.68 \frac{296}{(115+296)} = 0.75$ x 100% =75%

Total Calculation

Calculation This total calculation is the result of the sum of the Eligible Matrix, the Considered Matrix and the Unfit Matrix.
∑ T_pos = 373 Record
∑ F_pos = 38 Record
∑ T_neg = 784 Record
∑ F_neg = 38 Record

Sensivity (Total) $= \frac{t\_pos}{pos} = \frac{373}{411} = 0.90$

Specificity (Total) $= \frac{t\_neg}{neg} = \frac{784}{822} = 0.95$

Precision (Total) $= \frac{t\_pos}{(t\_pos + f\_pos)} = \frac{373}{(373+38)} = 0.90$

Accuracy (Total) = sensitivity $\frac{pos}{(pos + neg)}$ + specifity $\frac{neg}{(pos + neg)}$

$= 0.90 \frac{411}{(411+822)} + 0.95 \frac{822}{(411+822)} = 0.94$ x $\frac{1}{100}$ = 94 %

Calculation Average Accuracy

Accuracy (Eligible) x Accuracy (Considered) x Accuracy (Unfit)/3
= (97% + 89% + 75%) / 3 = 87%

Conclusion of Classifier Accuracy Measurables Evaluation, based on the evaluation of the Classifier Accuracy Measurables calculation, it is found that the accuracy of the predictions that have been made through the stage Knowledge Discovery in Databases (KDD)is 94%. The factors that determine the quality of the prediction are the True positive (T_pos) value: 373, False positive (F_pos): 38, True negative (T_neg): 784, and False negative (F_neg): 38, with the provision that the higher the T_pos and T_neg values, and the smaller the value of F_pos and F_neg, the greater the resulting accuracy value.

Conclusions

The application of KDD is used by staff from the Women Entrepreneurs event for the benefit of the executive so that the analysis process of Business Actor becomes easier because this analysis tool can extract large amounts of data and classify it into classes that are formed for potential contestants (Business Actors). The attributes used for class formation and knowing the profiles of prospective Competitors in this technique consist of four main attributes, namely Age Range, Length of Business, Type of Business 1 and Business Scale and five supporting attributes, namely Last Education, Occupation, Type of Business 2, Number of Personnel. Work and Online Marketing. The results of the Performance Evaluation of the Classification Method using the Classifier Accuracy Measurables method show that the accuracy of the predictions that have been made is 94%.

References

Cegielski, C., & Rainer, K. (2009). Introduction to Information Systems: Enabling and Transforming Business (pp. 622-634): New Jersey: John Wiley & Sons Inc.

Connolly, T. M., & Begg, C. E. (2005). *Database systems: a practical approach to design, implementation, and management*: Pearson Education.

Inmon, W. H. (2005). *Building the data warehouse*: John wiley & sons.

Işık, Ö., Jones, M. C., & Sidorova, A. (2013). Business intelligence success: The roles of BI capabilities and decision environments. *Information & management, 50*(1), 13-23. doi:https://doi.org/10.1016/j.im.2012.12.001

Kapoor, B. (2010). Business intelligence and its use for human resource management. *The Journal of Human Resource and Adult Learning, 6*(2), 21-30.

Kimball, R., & Ross, M. (2010). *The Kimball group reader: relentlessly practical tools for data warehousing and business intelligence*: John Wiley & Sons.

Laudon, K. C. (2007). *Management information systems: Managing the digital firm*: Pearson Education India.

Marakas, G. M., & O'Brien, J. A. (2013). *Introduction to information systems*: McGraw-Hill/Irwin New York, NY.

Pardillo, J., & Mazón, J.-N. (2011). Using ontologies for the design of data warehouses. *arXiv preprint arXiv:1106.0304*.

Person, R. (2013). *Balanced scorecards and operational dashboards with Microsoft Excel*: John Wiley & Sons.

Robbins, S. P., & Coulter, M. (2012). Management. England: Pearson Education Limited.

Turban, E., Sharda, R., Delen, D., & King, D. (2011). Business Intelligence. A Managerial Approach. 2. painos: New Jersey: Pearson Education Inc.

Vercellis, C. (2011). *Business intelligence: data mining and optimization for decision making*: John Wiley & Sons.