Discriminant Analysis of Sports Personalities for Predicting Better Performance

Namrata Bhawsar
Gyan Ganga Institute of Technology and Sciences, India

Santosh K. Vishwakarma
Gyan Ganga Institute of Technology and Sciences, India

ABSTRACT
Managing huge amount of data has always been a matter of concern. With the increasing growth in the sports domain, the amount of data in sports academy is also increasing. In this paper, Discriminant Analysis is used as a means of analyzing the effect of independent variables upon the dependent variables. Discriminant analysis is a method of data mining and it also has the ability to predict the category and class to which the focus belongs. With the help of this method, one can predict the performance of an athlete that can help the sportsperson and the trainer to choose their desired prime sport. The dataset used in the research is the battery test of athletes conducted in the sports academy. RapidMiner, an open source software is used to fetch results.

General Terms
Data Mining, Knowledge Discovery in Databases(KDD) Classification, Discriminant Analysis(DA), Performance Prediction.

Keywords
Data Mining, Knowledge Discovery in Databases(KDD), Linear Discriminant Analysis(LDA), Classification, Athlete's Performance Prediction.

1. INTRODUCTION
Data mining is the process of discovering interesting patterns from massive databases. Data is obtained from various sources including databases, data warehouses, web etc. It is a new powerful technology with great potential to help companies focus on most important information in their data warehouses. Knowledge discovery in databases(KDD), involves various steps like data selection, data integration, data cleaning, data transformation, pattern discovery, pattern evaluation and knowledge presentation[1]. Association, classification, clustering, prediction, sequential pattern and decision tree are the data mining techniques. Classification assigns objects in a group to focus on categories or classes. The main aim of classification is to predict the target class in the data. Prediction is same as classification. It use the existing variable present in the databases to predict the unknown variables. Discriminant Analysis is a method of data mining and it also has the ability to predict the class to which the target belongs. Discrimination is done by comparing the mean of the variables. The goal of this paper is to predict the performance of athletes to choose their desired prime sport. Prediction is done by considering the battery test conducted in the sports academy.

2. LITERATURE REVIEW
Over the past years, many research work has been done in this field. The 1st study titled Predicting Students Performance in University Courses: A Case Study and gear in KSU department of mathematics [3] constructs associate application to predict student's performance in Associate programming course supported their previous performances in specific arithmetic and English courses. To boot, the goal of design is to chop back left out rates by serving students, predict their performance in programming courses before their enrollment. Two experiments were conducted creating use of the CBA rule-generation rule. the primary used student's scores in math courses and in English courses, generating four rules with accuracy of 62.75%. The secondary used student's scores only in English courses, generating four rules with accuracy of 67.33%. The results of student's performance in English courses contains a vital clairvoyant impact on their performance inside the programming sources.

The 2nd study Predicting Consumer Purchase Intention A Discriminant Analysis Approach[6] Srivastava & Ali, 2013 found in their study that goodwill, friendly employees, proximity and specific product availability at the shop have completely different mean from the rest. Goodwill is that the most significant factor in choosing the outlet followed by standing, handiness of contemporary stock, stylish stock, promotional theme and shopping environment whereas proximity and therefore the handiness of a specific product at the shop area unit less influential factors in choosing a selected store. choice of the most well-liked attire store depends on the marital status of respondents and is freelance of their age, gender, variety of members within the family, education, employment standing and financial gain, frequency of searching and annual disbursal on the acquisition of apparel.

The 3rd study titled A Review on Predicting Students Performance using data mining Techniques [4] presents a summary on the data mining methods which are used to predict student's performance. This research study conjointly focuses on the prediction algorithmic program are often used to establish the foremost vital attributes in an exceedingly student's knowledge. The meta-analysis is accurate in prediction strategies and conjointly the most vital factors which will impact the student's performance. Classification technique is used to predict accuracy, classified by algorithms for predicting student's performance since 2002 to 2015. Neural Network has the most effective prediction accuracy by (98%) accompanied by decision tree (91%). Next, K Nearest Neighbor and Support Vector Machine gave a similar accuracy, that is (83%). At last, the tactic Naïve Bayes has lower prediction accuracy by (76%).

3. METHODOLOGY

Knowledge Discovery in Databases(KDD) is the process of transforming raw data into useful information. Data mining and KDD works simultaneously to extract information. KDD involves high iteration and interaction, starting with raw data and ending with fruitful information is the Knowledge Discovery in Databases. Following are the steps involved in KDD:

(1) Data Selection.

(2) Data Preprocessing.

(3) Data Transformation.

(4) Data Mining.

(5) Data Interpretation and Evaluation.

3.1 Data Selection

Data Selection is the process in which relevant data is retrieved from the database. The retrieved dataset are used for the research. Proper understanding of the application domain, prior knowledge related to the goals of ends users. The suitable kind of data is retrieved. Sports academy conducts several battery test of athletes in the field. These tests helps the athletes and their trainer to better understand their performance and choose desired best sport. A certain score is defined by the sports academy which must be achieved by the athletes. It is compulsory for every athlete to obtain that score based on which an athlete can be considered to be passed. The data set used in this research paper is the battery test of athletes conducted in the academy. This dataset provides information about athletes performance during the entire training period. The goal of this research paper is to predict the best sport in which they have secured best scores.

Age, Strength, Quickness, Injury, Vision, Endurance, Agility, Decision Making, and Prime Sport are the nine attributes through which the best sport for the athletes will be determined. Initially Prime Sport is a categorical attribute and .the remaining are numerical attributes. Description of each attribute is given below:

- Age- This is the age in years. Participants age ranged between 13-19.
- Strength- Measured through weight lifting exercises and recorded between 0-10, 0 being limited strength and 10 being sufficient.

	A	B	C	D	E	F	G	H	I	J	K
1	Age	Strength	Quickness	Injury	Vision	Endurance	Agility	Decision_M	Prime_Sport		
2	15	3	2	1	2	3	29	4	Football		
3	15	3	2	0	3	5	18	8	Baseball		
4	13	5	5	0	2	5	27	28	Hockey		
5	18	5	1	1	1	3	48	36	Hockey		
6	16	3	1	0	3	3	38	29	Football		
7	14	5	1	0	3	5	28	103	Basketball		
8	15	3	1	1	2	3	46	103	Football		
9	14	1	2	0	3	5	16	14	Football		
10	13	3	1	1	3	5	20	32	Football		
11	15	4	2	0	2	5	34	61	Basketball		
12	17	3	3	0	3	5	19	41	Baseball		
13	15	4	1	0	2	5	40	4	Hockey		
14	17	4	2	1	1	6	31	43	Football		
15	18	3	1	1	3	5	24	45	Basketball		

Fig. 1 sample of dataset

- Quickness- Rated between 0-6 with 6 being extremely quick and 0 being very slow.
- Injury- This is simple yes/no, common injury which can be treated with ice, rest, stretching were entered as 0. Injuries that took more than three weeks to heal or require some physical surgery were flagged as 1.
- Vision- Tested on 20/20 vision scale using eyechart. Athletes scored between 0-4 scale, 4 with being perfect.
- Endurance- Physical fitness test including running, aerobics and cardiovascular exercises and distance
- swimming. Their performance was rated between 0-10.
- Agility- series of test for their ability to move, twist, turn, jump, change direction. Athletes scored between 0-100.
- Decision Making- tests the athlete's process of deciding what to do in any athletic situation. Scores recorded between 0-100.

3.2 Preprocessing and Cleaning of Data

This step is two-folded. Firstly the data from multiple heterogeneous sources is merged into a coherent data store. Data processing is the process of converting data into more readable and understandable format. Raw data is of no use, quality of data is the prime affair of datamining. Data acquired from various sources may contain inconsistent, reductant and noisy data, which may results in low quality and inappropriate data. So preprocessing of data is essential to enhance the quality and efficiency of the dataset. This step provides us useful and efficient data. This is called Preprocessing and it results in inconsistent data. Secondly, to remove the noisy data and correct the inconsistencies some transformations are applied. This is called Data Cleaning.

- Data cleaning is the process of removing noise, inconsistent data and missing value from the

dataset. Missing value of the attributes are replaced by some corresponding values. In our dataset, some inconsistent values are present that are removed.

- All the values are not important in dataset for datamining. So the deduction of such data is necessary to make dataset suitable without the information loss. Those values which are important for the research are only present in dataset.

ExampleSet (493 examples, 0 special attributes, 9 regular attributes)

Row No.	Age	Strength	Quickness	Injury	Vision	Endurance	Agility	Decision... ↑	Prime_Sport
400	15	3	1	0	2	3	35	0	Hockey
307	13	3	1	0	1	3	40	1	Hockey
301	18	3	2	1	3	5	20	2	Baseball
72	18	6	1	1	2	1	53	3	Basketball
202	18	3	1	0	2	3	38	3	Football
274	17	1	1	0	3	5	17	3	Baseball
432	18	4	2	1	2	5	25	3	Hockey
1	15	3	2	1	2	3	29	4	Football
12	15	4	1	0	2	5	40	4	Hockey
17	14	3	0	1	0	3	27	4	Baseball
39	13	3	1	0	3	5	20	4	Basketball
45	18	5	1	1	2	3	46	4	Football
55	18	4	2	1	0	3	32	4	Hockey
65	15	4	5	0	0	3	23	4	Hockey

Fig. 2 Pre Processed dataset

3.3 Feature Selection and Extraction

Feature selection is the process of selecting a subset of original features or attributes. The underlying idea is that the original data contains many attributes which are irrelevant/redundant and hence can be removed without incurring significant loss of information. Feature extraction is the process of applying transformations to the attributes in the original data to project new features which reduces the number of attributes and thereby reducing dimensionality. Data transformation is the process of eliminating the number of values. Transformed data is shown in figure 3.Certain operators are used to remove the values which are not appropriate for the research. Rather than using bad data, we will simply remove them.

- If you will see Decision Making attribute in original dataset, you will find some values which are not important for the research like values smaller than3. Those athletes who got less than 3 in Decision Making attribute will be removed from the dataset. Use Filter Example operator in the main process and set the Condition Class to attribute_ value_ filter and enter Decision_ Making>=3 in the Parameter string. This will remove all the values which are smaller than three.

- Similarly, the same process will repeat and those athletes whose Decision_ Making attribute is greater than 100 will also be removed from dataset. Set the Condition Class to attribute_ value_ filter and enter Decision_ Making<=100 in the Parameter string. This will remove all the values which are greater than 100 from the dataset. After this step number of observation has been reduced to 482 from 494 observations.

ExampleSet (482 examples, 0 special attributes, 9 regular attributes)

Row No.	Age	Strength	Quickness	Injury	Vision	Endurance	Agility	Decision_M...	Prime_Sport
1	15	3	2	1	2	3	29	4	Football
2	15	3	2	0	3	5	18	8	Baseball
3	13	5	5	0	2	5	27	28	Hockey
4	18	5	1	1	1	3	48	36	Hockey
5	16	3	1	0	3	3	38	29	Football
6	14	1	2	0	3	5	16	14	Football
7	13	3	1	1	3	5	20	32	Football
8	15	4	2	0	2	5	34	61	Basketball
9	17	3	3	0	3	5	19	41	Baseball
10	15	4	1	0	2	5	40	4	Hockey
11	17	4	2	1	1	6	31	43	Football
12	18	3	1	1	3	5	24	45	Basketball
13	17	4	2	0	0	5	24	32	Basketball
14	18	5	2	0	0	5	47	35	Football
15	14	3	0	1	0	3	27	4	Baseball

Fig. 3 Transformed dataset

3.4 Data Mining

Data mining is the process of analyzing data from different dimensions and angles and summarizing it into useful and sensible structure for further use. It is the most important step in KDD, useful information is extracted from the transformed data. Classification is the technique used for data mining to obtain suitable outcomes. Discriminant Analysis is a method of data mining and it also has the ability to predict the category or class to which the focus belongs. This method aims to develop discriminant function which is just a combination of independent variable that will discriminant between the classes of the dependent variables. A dependent variable is the value which the examiner wants to predict. Discriminant Analysis is a statistical tool to classify individuals into groups or categories based on certain independent variables. This is done by developing Discriminant functions that are linear combination of independent variables that will discriminate between the defined categories.

When there are two categories for classification, then the analysis is termed as Two-Group Discriminant Analysis. If there are three or more than three categories, then the analysis is called Multiple Discriminant Analysis. The major distinction between the two types is that for two groups, it is possible to derive only one Discriminant function whereas for multiple Discriminant analysis more than one Discriminant function can be computed.

Below the steps involved in Discriminant analysis:

- Problem Formulation
- Estimation of Discriminant Function Coefficients
- Determination of significance of Discriminant
- Interpretation of results
- Validity Assessment

For example, we can classify high school graduates into two groups: Those who choose to attend college after graduation and those who do not. The independent variable could be the intention of student to continue to college which could have been surveyed one ear prior to graduation. Then the mean of the two groups can be calculated. If the means for the two groups (those who went to college and those who did not) are different, then we can say that intention to attend college as stated one year prior to graduation allows us to discriminate between those who are and are not going to attend college after graduation.

Stepwise Discriminant Analysis

Stepwise Discriminant analysis means step by step inclusion or exclusion of such independent variables which contribute most to the discrimination between groups. There are two ways to conduct stepwise Discriminant analysis:

Forward stepwise analysis : In Forward Stepwise Analysis the discrimination model is built step by step. At each step we review and evaluate all independent variables to determine the ones which will contribute most to the model. The variable will then be included and the process will start again.

Backward stepwise analysis : In this case all the variables are initially included in the model and then at each step , the variables that contribute least to the determination of group classification are eliminated. Thus in the end only the most beneficial independent variables remain in the model.

Result of Discriminant Analysis on our training dataset.

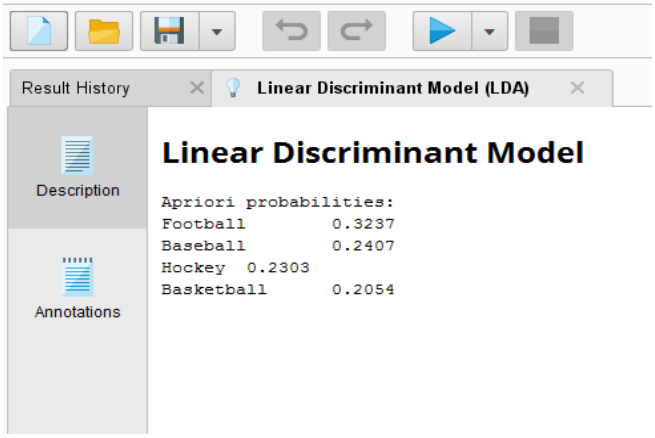

Fig. 4 result of Discriminant Analysis

Sum of all the probabilities given are equal to 1. Probabilities can be calculated very easily with RapidMiner. Probability of football is 0.3237, if you will look back at the figure-----, football as Prime Sport covers 156 of our 482 observations. So the probability of football is 156/482=0.3237. Like this all the probabilities will be calculated.

3.5 Interpretation and Evaluation

In this step, the mined data Patterns are evaluated and interpreted with respect to the defined goals. The focus here is the comprehensibility and usefulness of the whole model. The mined knowledge is also documented for further usage in this step. By applying the data mining technique (classification) a pattern is attained from the transformed data, in other words training of dataset is done.

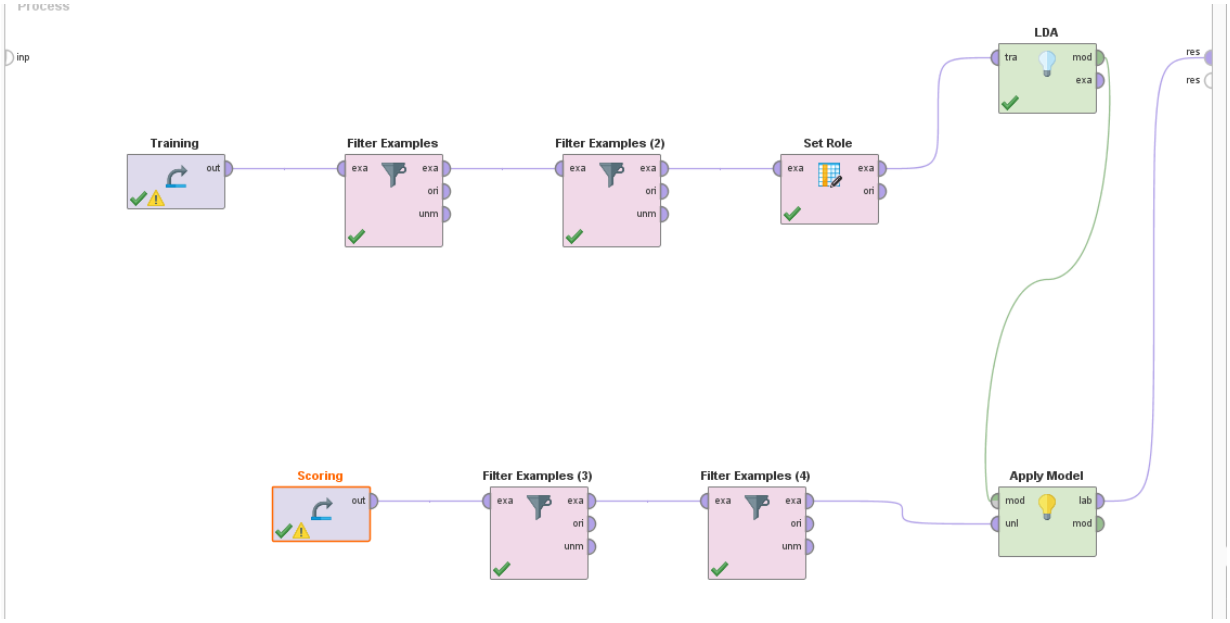

Fig. 5 Linear Discriminant Model

Take a different dataset in main process window and rename it as scoring. Inconsistent data present in Decision Making attribute is removed from the same process as done before. Add two Filter Example operator and set the Condition Class to attribute_ value_ filter and enter Decision_ Making>=3 in 1st filter example and Decision_ Making<=100 in the 2nd filter example respectively. To complete our model and predict the Prime_Sport for 1767 boys present in our Scoring data set.

We need an operator called as Apply Model. Drag and drop Apply Model and connect it with Filter Example. This will give error because Apply Model operator expects the output of a model generation operator as its input. LDA operator has a mod port for its output. So simply connect the mod port of the LDA operator with the mod port of the Apply Model. Apply model operator applies the training data model on the scoring data for prediction. In our research, we want to predict the Prime Sport for the athletes of the academy based on the previous performances.

ExampleSet (1767 examples, 5 special attributes, 8 regular attributes)

Row No.	prediction(P...	confid...	confide...	confide...	confide...	Age	Strength	Quickness	Injury	Vision	Endurance	Agility	Decision_Making
1	Basketball	0	0	0	0	18	5	1	1	0	5	33	61
2	Baseball	0	0	0	0	13	1	2	1	3	5	18	59
3	Football	0	0	0	0	13	2	1	0	2	5	40	11
4	Baseball	0	0	0	0	16	3	1	0	2	5	32	35
5	Football	0	0	0	0	15	1	1	0	2	3	43	37
6	Baseball	0	0	0	0	17	3	2	0	3	5	21	41
7	Football	0	0	0	0	16	3	1	0	1	1	41	29
8	Baseball	0	0	0	0	15	1	2	1	3	5	17	45
9	Football	0	0	0	0	16	3	2	0	1	3	46	40
10	Football	0	0	0	0	18	5	1	1	2	5	41	6
11	Football	0	0	0	0	14	5	2	1	0	3	35	48
12	Baseball	0	0	0	0	17	2	2	1	2	1	28	32
13	Football	0	0	0	0	14	6	1	1	3	3	42	39
14	Football	0	0	0	0	15	4	1	1	1	1	49	4
15	Basketball	0	0	0	0	14	5	1	1	0	5	24	55
16	Hockey	0	0	0	0	14	4	2	1	0	5	21	45
17	Football	0	0	0	0	13	5	1	1	3	3	42	29
18	Hockey	0	0	0	0	13	4	2	1	0	3	27	7
19	Baseball	0	0	0	0	17	0	0	0	3	5	15	43
20	Football	0	0	0	0	15	5	1	0	2	5	31	4
21	Football	0	0	0	0	17	2	5	1	2	1	42	6
22	Baseball	0	0	0	0	13	3	1	1	3	5	21	29
23	Basketball	0	0	0	0	16	5	3	1	0	5	38	75

Fig. 6 prediction generated by rapidminer

Prime Sport is the prediction result for each boy of the academy based on the specialization sport of the former academy students.

4. CONCLUSION AND FUTURE WORK

This research makes use of Discriminant Analysis to predict the best sport for the athletes based on their battery test and the former athletes performance. DA helps us go across the edge of classification and prediction in datamining. discriminant analysis is similar to k-means clustering. Some application of DA is to help worker, students, athletes, organization and many more to predict their potentially successful paths based on their characterization. For future work, we might prefer to refine our work by taking additional variety of example set and are available up with additional accuracy and alternative techniques to assist students in their academic careers.

5. REFERENCES

[1] Han, J., Kamber, M., Pei, J.. Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann; 3rd ed.; 2012. ISBN 978-0-12- 381479-1.

[2] Pujari, K., A.. Data Mining Techniques. ISBN-10:8173713804.

[3] Badr, G., Algobail, A., Almutairi, H., Almutery, M., Predicting Students Performance in University Courses: A Case Study and Tool in KSU Mathematics Department. Procedia Computer Science 82 (2016) 80 89.

[4] Shahiri,A., M., Husain,W., Rashid,N.,A., A Review on Predicting Students Performance using Data Mining Techniques. Procedia Computer Science 72 (2015) 414 422.

[5] RapidminerStudioDocumentation http://docs.rapidminer.com/studio/

[6] Predicting Consumer Purchase Intention: A Discriminant Analysis Approach Sougata Banerjee Sarwat Pawar

[7] www.google.co.in/