Volume-9,Issue-2(Feb-18)
ISSN (O) :- 2349-3585

A CASE STUDY ANALYSIS FOR STUDENTS NOT PARTICIPATE IN SPORTS USING DATA MINING TECHNIQUES

Ms. Amala[1],B. Jasmine[2],Ms. Karthiga[3]
[1,3]PG Student,[2]Asst. Professor, Jayaraj Annapackiam College for women (Autonomous), Periyakulam,

ABSTRACT:This study was aimed at finding reasons for non-participation in sport by at school and College level in Theni District. According to the findings of this research, factors that have the most important influence on non-participation in sport by students are (in order of importance), facilities, political factors, social factors, self-image, economic factors and health. In terms of facilities it has been revealed that black township schools do not have adequate equipment, properly organized recreational facilities, coaches for the different sport codes and upgraded as well as well-maintained sport fields. Cultural isolation of black players in sport surfaced as an important political factor for non-participation. Through this research it has also emerged that while gender and income do seem to have an influence on non-participation in sport, grades and home environment do not. The results indicate that income as a reason for non-participation in sport is significantly more important for learners from low-income families than for learners from average-income families.

Key words: Data Mining, K-Means Cluster.

1. Introduction:

Sports consist of physical and mentally competitive activities carried out with a Recreational purpose for competition, for self-enjoyment, to attain excellence, for the Development of a skill, or some combination of these (en.wikipedia.org). Sport was also defined by as "an activity which offers the individual the opportunity of self-knowledge, self-expression and fulfillment; personal achievement, skill acquisition and demonstration of ability; social integration, enjoyment, good health and well-being." Physical activity on the other hand is defined by as "bodily movement produced by the contraction of skeletal muscle that substantially increases energy expenditure above the basal level. Common categories include occupational, household, leisure-time, (including competitive sports, recreational activities, exercise training) or transportation."

1.2. Present Condition:

This study examined perceived athletic identity, sport commitment, and the effect of sport participation to identify the impact of athletic participation on college students. This study surveyed 163 student-athletes (59%) and 112 non-athlete students (41%) from a National Collegiate Athletic Association Division-I affiliated institution (males = 172, 62.5%; females = 103, 37.5%). The survey questionnaire was developed and modified from four well-established instruments, the Athletic Identity Measurement Scale, the Sport Commitment Model, the Life Roles Inventory-Values Scales, and Athletic Involvement on the Social Life. The data collection process was initiated and completed in the 2008 spring semester.

2. Algorithm used

2.1. Cluster Analysis:

Cluster analysis is a multivariate analysis that attempts to form groups or "clusters" of objects (Sample Plots in our case) that are "similar" to each other but which differ among clusters. The exact definition of "similar" is variable among algorithms. But has a generic basis. The methods of forming clusters also vary, but follow a few general blueprints.

2.2. Similarity, Dissimilarity and Distance:

Similarity is characterization of the ratio of the number of attributes two objects share in common compared to the total list of attributes between them. Objects which have everything in common are identical, and have a similarity of 1.0. Objects which have nothing in common have a similarity of 0.0. As we have discussed previously, there is a large number of similarity indices proposed and employed, but the concepts are common to all.Dissimilarity is the complement of similarity and is a characterization of the number of attributes two objects have uniquely compared to the total list of attribute between them. In general, dissimilarity can be calculated as 1-similarity.

2.3. K-means clustering

The most common partitioning method is the K-means cluster analysis.

Conceptually, the K-means algorithm:

1. Selects K cancroids (K rows chosen at random)
2. Assigns each data point to its closest centroid
3. Recaculates the centroids as the average of all data points in a cluster(i.e.,thecentriods are p-length mean vectors, where p is the number of variables)
4. Assigns data points to their closer set centroids
5. Continues step 3 and 4 until the observations are not reassigned or the maximum number of iterations(R uses 10 as a default) is reached.

3. TOOLS FOR THE STUDY

3.1 The R Environment:

R is free software environment for statistical computing and graphics. It provides a wide variety of statistical and graphics techniques. R can be extended easily via packages. R is an integrated suite of software facilities for data manipulation, calculation and graphics display.

Cluster Analysis in R

R has an amazing variety of function for cluster analysis. In this section, We use three of the many approaches: hierarchical agglomerative, partitioning, and model base

Data preparation: Prior to clustering data, you may want to remove or estimate missing data and rescale variables for comparability.

```
#prepare Data
        Mydata<-na.omit(mydata) # listwise deletion of missing
        Mydata<- scale(mydata)
```

Partitioning: K-means clustering is the most popular partitioning methods. It requires the analyst to specify the number of cluster to extract. A plot of the within groups sum of squares by number of cluster extracted can help determine the appropriate number of cluster. The analyst looks for a bend in the plot similar to a screen test in factor analysis.

```
# Determine number of cluster
        >wss<- (nrow(mydata)-1)*sum(apply(mydata,2,var))
        >for (i in 2:27) wss[i] <- sum(kmeans(mydata,centers=i)$withinss)
        >plot(1:27, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares")
#K-means cluster analysis
        > fit <- kmeans(mydata, 5)
# get Cluster means
        >aggregate(mydata,by=list(fit$cluster),FUN=mean)
# append cluster assignment
        Mydata<- data.frame(mydata, fit$cluster)
```

A robust version of K-means based on mediods can be invoked by using pam() instead of kmeans(). The function pamk() in the fpc package is a wrapper for pam that also prints the suggested number of cluster based on optimum average silhouette width

3.2. Statistical techniques used

Data sources and methodology

Target population:

This survey covers all the students of Theni district. **Instrument design:** This questionnaire collects data on the attitude of the school and college students. The items and reasons on the questionnaire have remained unchanged for several years. Howeve r, should modifications become necessary, proposed changes would go through a review committee and a field test with respondents and data users to ensure its relevancy.

Sampling: This survey is a census with a cross-sectional design. Data are collected for particular units of the target population, therefore sampling is done.

Data sources: Responding to this, survey is mandatory. Data are collected directly from survey respondents. Data are compiled from the responses the researcher collected by the questionnaire. The researcher performs the data capture activities, and follow-up of non-respondents. Contact with respondents is maintained for subsequent follow-up.

Error detection: There are edits built into the data capture application to check the entered data for unusual values, as well as to check for logical inconsistencies. Whenever an edit fails, the interviewer is prompted to correct the information (with the help of the respondent when necessary). For most edit failures the interviewer has the ability to override the edit failure if necessary.

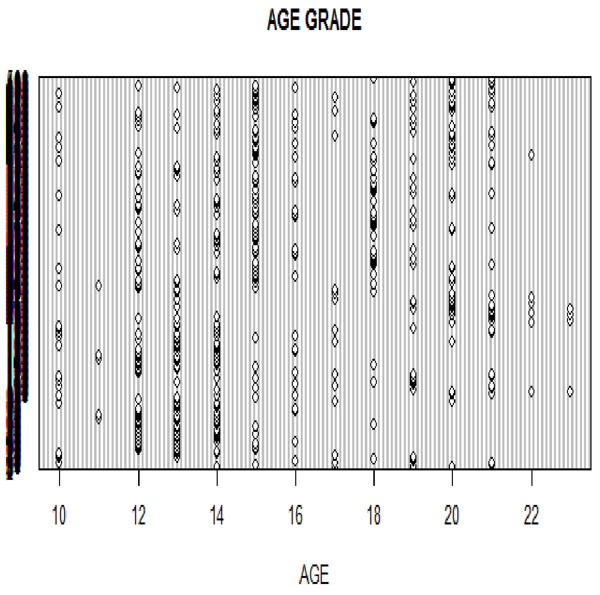
Imputation: A 100% response rate is attained; therefore imputation is not necessary.

Quality evaluation: Prior to the data release, combined survey results are analyzed for comparability; in general, this includes a detailed review of individual responses, general economic conditions, and historical trends. The data is examined at a macro level to ensure that the long-term trends make sense when compared to publicly available information in media reports, and etc.

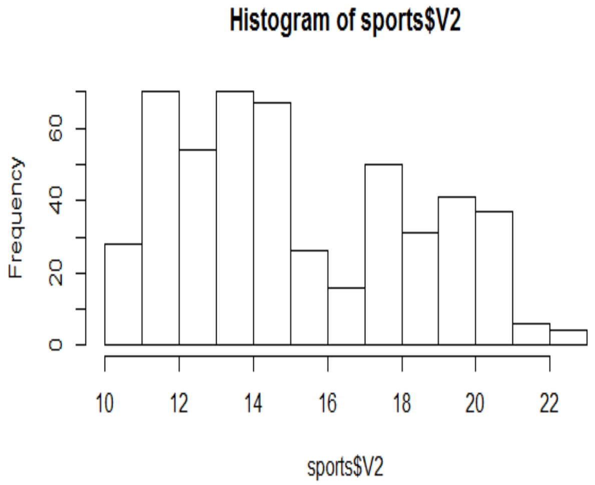
Revisions and seasonal adjustment: Revisions in the raw data are required to correct known non-sampling errors. These normally include replacing imputed data with reported data, corrections to previously reported data, and estimates for new births that were not known at the time of the original estimates. Raw data are revised, on a monthly basis, for the month immediately prior to the current reference month being published. The purpose is to correct any significant problems that have been found that apply for an extended period. The actual period of revision depends on the nature of the problem identified.

4. FINDINGS AND INTERPRETATIONS

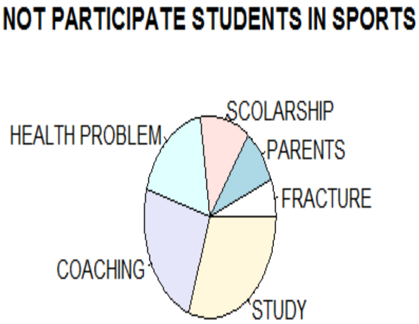
AGE BASED:



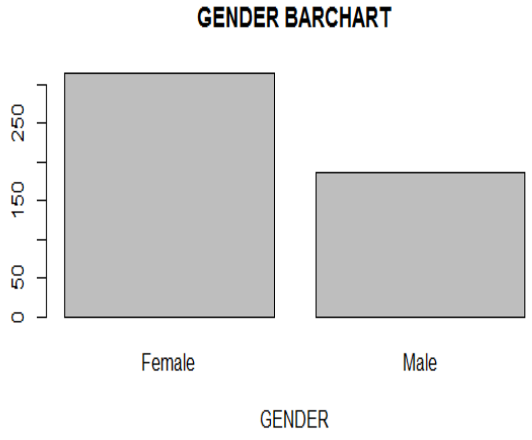
HISTOGRAM:



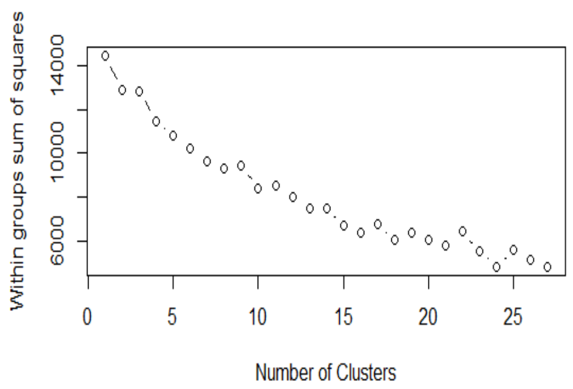
PIECHART:



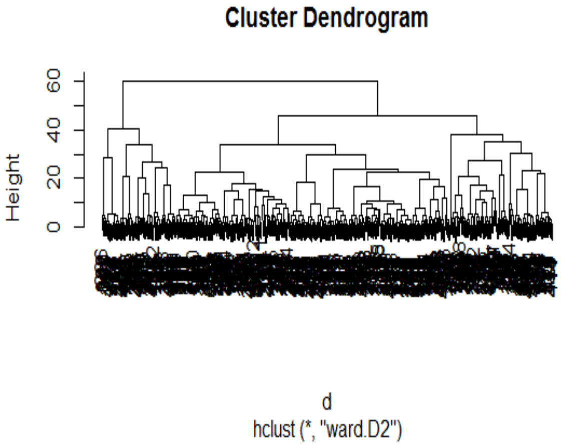
GENDER BASED BARPLOT:



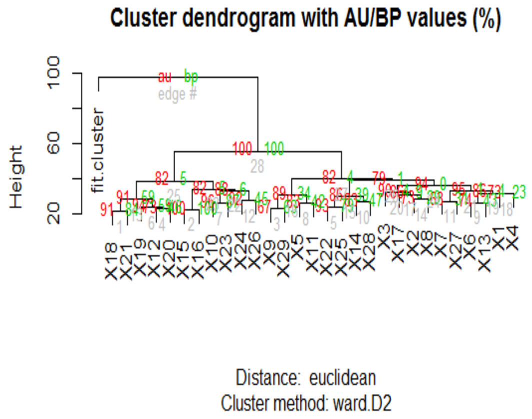
```
>for (i in 2:27) wss[i] <- sum(kmeans(mydata,centers=i)$withinss)
>plot(1:27, wss, type="b", xlab="Number of Clusters",
ylab="Within groups sum of squares")
```



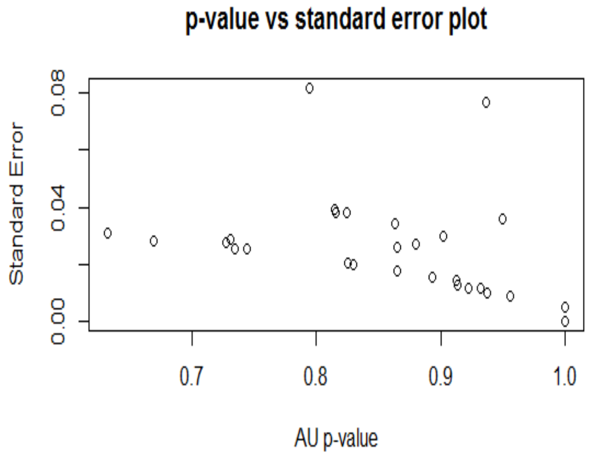
```
>mydata<- data.frame(mydata, fit$cluster)
> d <- dist(mydata, method = "euclidean")
> fit <- hclust(d, method="ward.D2")
>plot(fit)
```



```
>plot(fit)
```



```
> x <- seplot(fit, identify=TRUE)
```



Model Based:

```
>Msplot(fit, edges=2)
>library(mclust)
> fit <- Mclust(mydata)
>plot(fit)
```

Model-based clustering plots:

1: BIC

2: classification

3: uncertainty

4: density

Selection: 1

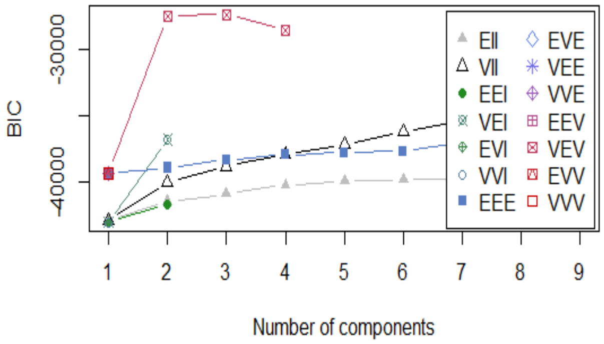
Model-based clustering plots:

1: BIC

2: classification

3: uncertainty

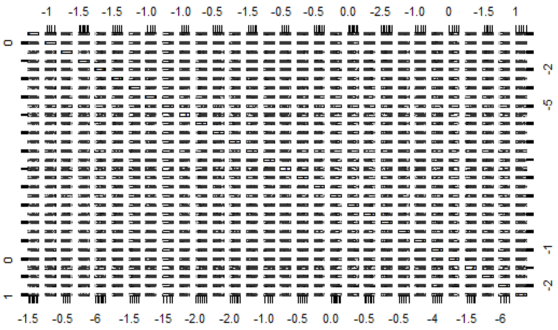
4: density



Selection: 2

Model-based clustering plots:

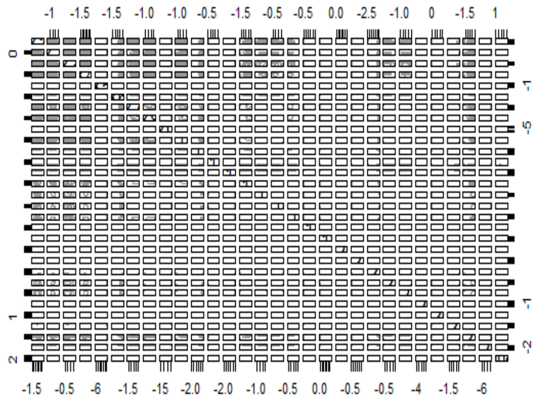
```
1: BIC
2: classification
3: uncertainty
4: density
```



Selection: 3

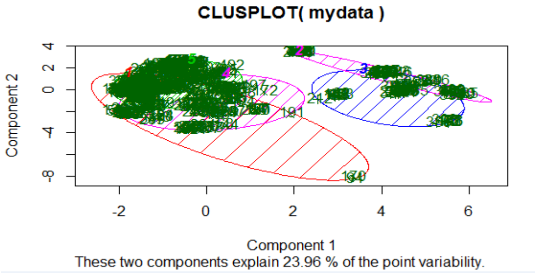
Model-based clustering plots:

```
1: BIC
2: classification
3: uncertainty
4: density
```

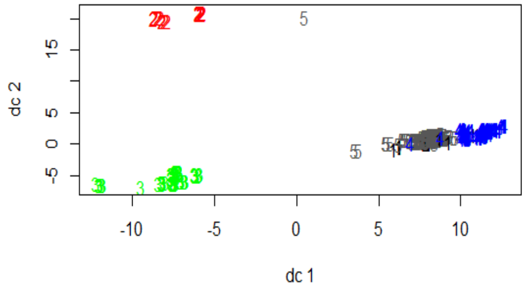


Plotting Cluster:

```
> fit <- kmeans(mydata, 5)
>library(cluster)
>clusplot(mydata, fit$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)
```



```
>library(fpc)
>plotcluster(mydata,fit$cluster)
```



5. RECOMMENDATIONS AND SUGGESTIONS FOR FURTHER RESEARCH

5.1. RECOMMENDATIONS

Following this study a number of recommendations can be made:This research was done using only schools and college in Theni district. Government, Department of Education and Policy Planners must read this and related reports. The main factors responsible for nonparticipation identified by this study are facilities, political and social which are all in the sociopolitical realm and are amenable to redress through the sociopolitical process. Greater budgetary allocations, stricter husbandry of monies allocated and greater deployment of appropriate personal would go a long way to alleviating the situation.The government must introduce properly thought out and coordinated policies which encompass sporting activity at every level in the school system. These policies should be backed up with appropriate management and budget.The significance of appropriate training of coaches should not be overlooked. Such training should include a variety of topics to address diversity issues, racial bias and cultural separation of the players that will lead to the provision of a positive sport environment. Coaches should also be well informed about how to motivate learners to participate and adhere in sport and about the maintenance of a good relationship between the coach, parents and learners. It should be instilled in them that for successful and lasting participation of learners, fun and enjoyment should be the fundamental elements in sports.

5.2. SUGGESTIONS FOR FURTHER RESEARCH

There are several promising directions to extend the work presented is this case study:

- Find the effects of different types of problems on student achievement.
- Develop techniques that apply student information in helping individulas of user source more efficiently
- Identify those students who are at risk of failure, especially in very large groups.

6. BIBILIOGRAPHY

6.1 References :

[1]. Alegi, P C 2004. UmdlaloWabantu: A History of Soccer in Pre-Apartheid South Africa. Boston University: Unpublished Document.

[2]. Allen, J B 2003. Social Motivation in Youth Sport.Journal of Sport and Exercise Psychology, 25(4):551-567.

[3].American Heritage Dictionary of the English Language. Fourth Edition 2000 Boston: Houghton Mifflin.

[4]. Andersen, M B 2000. Doing Sport Psychology.Melbourne: Human Kinetics.

[5]. Antshel, K M &Anderman, E M 2000. Social Influences on Sports Participation During Adolescence. Journal of Research and Development in Education, 33(2); 85-94.