International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 12 (2017) pp. 3298-3304
© Research India Publications. http://www.ripublication.com

Experience in Applying Data Mining Techniques to Musical Content Database to Identify Personality Traits

[1]Xavier Campaña, [1]Rubén Arroyo and [1,2,*]Sang Guun Yoo

Departamento de Ciencias de la Computación, Universidad de las Fuerzas Armadas ESPE, Av. General Rumiñahui s/n, Sangolquí, Ecuador.

Facultad de Ingeniería de Sistemas, Escuela Politécnica Nacional, Ladrón de Guevara, E11-253, Quito, Ecuador.

ORCID ID: 0000-0003-1376-3843, Scopus Author ID: 36187649600, Researcher ID: R-5327-2016

Abstract

The present work takes a tour about the influence of music in the society and tries to find the connections between the personality and the musical preferences in people. The work is proposed to identify a mechanism capable of evaluating personality; the mechanism is framed within five major factors embodied in the OCEAN model, which was determined many years ago by previous studies in Psychology. It includes a review of the different techniques used in the data mining methodology to find the associations between musical preferences and personality features and then it leads to the possible associations of both aspects.

Keywords: OCEAN model, data mining, personality, music

INTRODUCTION

Samples of musical expression began to appear during the evolution of the early hominids, both based on vocal sounds and musical instruments of wood, bone or stone. The process of musical production became more complex and the meaning of music became more important and it became part of activities, rites and artistic and emotional expressions [1].

The human brain has developed to analyze and produce music, regardless of the musical education of the individual [2] and because of this, music is reflected as the main manifestation of expression, identification and personality. All musical genres have a huge influence in our personality (at the individual level) and in our idiosyncrasy (at the level of society). Therefore, we can say that music has been an essential form of identification and expression people from the early ages of the hominids.

In the psychology field, there is a study scheme that examines the structure of personality based on five broad elements or personality traits (personality dimensions) [3] and it is considered one of the most recognized models. The five major personality traits, also called major factors, are often given the following names: factor O (openness to new experiences), factor C (consciousness), factor E (extroversion), factor A (agreeableness) and factor N (neuroticism or emotional instability), which forms the acronym "OCEAN" [4-5]. It is important to mention that each of the five major personality traits is constituted by a set of more specific personality traits; for example, the E factor (extroversion) includes concrete qualities such as the search for emotions, sociability or optimism.

This works seeks to exploit a database with musical tastes of people using the OCEAN model and it will combine with data mining techniques to extract important conclusions i.e. significant correlations between personality traits and musical preferences of the individuals.

BACKGROUND

Fundamental Concepts

Personality is the set of physical, genetic and social characteristics that make up an individual unique. The interrelation and communion of the mentioned elements generally generates the characteristics which will determine the behavior of a person [6]. The personality is composed of two elements: temperament and character, one has a genetic origin and the other is influenced by the social environment where the individual lives. In psychology, the Model of the Big Five is used to analyze the personality as the composition of five broad factors or dimensions of personality [4] and it was used on various methodologies to evaluate the five traits in an individual [7].

Music is, according to its traditional definition, the art of organizing a coherent combination of sounds and silences in sensible and logical way using the fundamental principles of melody, harmony and rhythm, through the intervention of complex psycho-psychic processes. The purpose of this art is to arouse an aesthetic experience in the listener, and to express feelings, emotions, circumstances, thoughts or ideas. Music is a stimulus that affects the perceptual field of the individual; thus, the sound flow can fulfill several functions such as entertainment and communication [2].

Music can explore and penetrate an individual's emotional awareness, transferring a series of emotions and feelings that are part of his/her sonorous-musical history. The sound file that each individual stores generates an identity as an autonomous being and as part of a social group involving the individual's culture. It means that the music becomes the individual's emotional language. Emotional language is comprehensible to human beings at a universal level, and therefore, it becomes a constructive element of the personality of the individual [1].

Data mining or data exploration (the Knowledge Discovery in Databases or KDD) is a field of computer science concerned with the process that attempts to discover patterns in large volume of data sets. It uses the methods of artificial intelligence, automatic learning, statistics and database systems. The goal of this technique is to find patterns that are previously unknown. Once these patterns are found they can be used to make certain decisions [8].

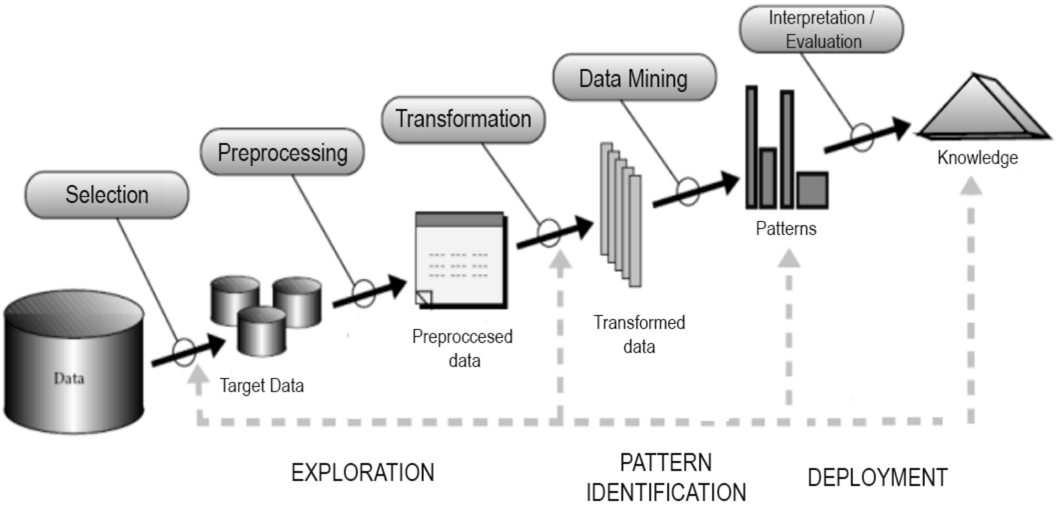


Figure 1: Process of discovery of knowledge

Fig. 1 shows the three general steps of the process of discovery of knowledge:

Exploration. In the first step, the data is cleaned and transformed. The important variables and the nature of the data are determined based on the data.

Identification of patterns. Once the data has been exploited, refined, and defined, patterns that offer the best prediction are identified and chosen.

Deployment. Finally in this step, the patterns are used to get the desired results.

Several algorithms and techniques are used for the discovery of knowledge from databases, such as:

Classification Algorithm: It uses a set of pre-classified examples to develop a model that can classify the population of records in general. The data classification process involves learning and classification. In learning, training data are analyzed using a classification algorithm. In the classification stage, the testing data is used to estimate the accuracy of the classification rules. If accuracy is acceptable, rules can be applied to the new data. This type of algorithm has been used in the development of decision Trees and Bayesian classifiers.

Grouping Algorithm: It identifies similar classes of data. By using clustering techniques, the density of regions can be identified in the object's space, as well as the general distribution pattern and the correlations between the attributes of the data. In this work, the "K-means" clustering algorithm has been used.

Neural Network Algorithm: The neural network is a set of connected input/output units with weighted links. During the learning phase, the network learns by adjusting the weights of each connection to allow the network to predict the correct class labels of the logs or input data. Neural networks have the ability to derive meaning from complicated or imprecise data and it can be used to extract patterns and detect trends that are too complex to be noticed by humans. In this work, the perceptron neural network algorithm has been used.

State of the art

There are several scientific studies that have already explored the correlation between music and personality. Some published papers regarding the subject of study are: "Music preference correlates of Jungian types" whose purpose was to explore the relationship between personality preferences and music, using the Myers Briggs Indicator [9]. In another study entitled: "Music Preference and the Five-Factor Model of the NEO Personality Inventory", the authors used a questionnaire that measures musical preference [10]. Other study called "Adolescents' music preferences and personality characteristics" examines the structure of musical preferences of Dutch adolescents, the stability of musical preferences, the relationships between the Five Great personality characteristics and changes in musical preferences [11]. Another paper published in the "Journal of Youth and Adolescence" titled "Music Preferences, Personality Style, and Developmental Issues of Adolescents" aimed at examining the personality characteristics and developmental problems of the 3 groups of adolescent music listeners: Those who prefer contemporary music, those who prefer heavy music, and those who have preferences or selected pieces only [12]. In addition to the previous mentioned studies, Samuel Gosling and Peter Rentfrow, from the University of Texas, distinguish 4 categories in which to group the musical context and in which the gross population is identified. Thus, followers of classical music, blues, jazz and soul are characterized by their emotional solidity and, as claimed by Cambridge researchers, are usually tolerant and open minded. Country and pop fans are revealed as more conservative and outgoing individuals. They enjoy an intense social life and the agglomerations of people but curiously, their verbal abilities are not too developed. Rock and heavy metal are identified with a degree of rebelliousness and impulsivity that makes them stand out wherever they go. They have a personal style and rely firmly on their intelligence. Lovers of soul, funk and hip-hop, as well as electronic music, are open, liberal and somewhat clueless. They are fascinated by sport and with certain exceptions, they do not pay much attention to the faults of others [13].

Even though, there are many studies that relate music to personalities, there is still lack of research done using an open database with data mining techniques. This is because data mining field has just begun to exploit and disseminate the use of both audio and video streaming platforms for the current connection facilities.

METHODOLOGY AND TOOLS

The purpose of the following section is to explain the methodologies and tools used in this research work. In this study, descriptive correlational research was used to determine the degree of non-causal relationship or association between two or more variables. First, the variables of Personality Inventory NEO-PI-R (see section 4.2) are measured, and then, through tests of correlational hypothesis and application of statistical techniques such as SEMMA Methodology (see section 4.4), the existing correlation is estimated [14].

The following subsections explain more about the SEMMA methodology and the Personality Inventories that were tools used for this research.

Data Mining Methodology: SEMMA

SAS Institute which is the developer of the SEMMA methodology [15] defines it as the process of selection, exploration and modeling of data to discover unknown business patterns. The methodology is comprised of the following phases.

1. Sampling. It identifies a significant representation of the population to be objective of the analysis. This is done with the aim of facilitating the mining processes on the data, reducing the time that it is necessary to determine the valuable information for the business.
2. Exploration. In this phase, the data extracted in the sample is reviewed to detect, identify and eliminate anomalous data, helping to refine the processes of information discovery in later phases of the process.
3. Modification. This is the process of creation, selection and transformation of the data or variables from which the model will be based on.
4. Modeling. This step makes use of software tools and data mining techniques/algorithms in order to obtain the desired information. The choice of model will depend essentially on the available data and variable types.
5. Assessment. Finally an analysis of the results is performed. The results obtained from the models are compared between each other.

Personality Inventories

There are different tools in the field of psychology aimed to evaluate and understand the personality of the human being. The rating scales and checklists have been designed for psych diagnosis and clinical purposes in educational contexts, identification of emotional disorders, and in the diagnosis of psychopathology in clinical situations. Personality inventories consist of a list of questions related to personal characteristics of individuals (thoughts, feelings and behavior). The veracity and validity of the personal inventories can be compromised if they are not answered truthfully, consciously and consistently. Goldberg [4] and Costa and McRae [5] defined the five personality factors, which appear to be very consistent among several groups of people and in different situations. The personality factors in turn include facets prone to being evaluated in the same way as the personal inventories [16]. These indicators are described in Table 1.

Table 1: Neo Personality Inventory Factors and Evaluations

Factor	Scales
(N) Neuroticism vs Emotional Stability	N1. Anxiety N2. Hostility N3. Depression N4. Social Anxiety N5.Impulsiveness N6.Vulnerability
(E) Extroversion	E1. Politeness E2.Gregariousness E3.Assertiveness E4.Activity E5.Search for emotions E6.Passive emotions
(O) Openness to new experiences	O1.Fantasy O2.Aesthetics O3.Feelings O4. Actions O5. Ideas O6. Values
(A) Affability or Courtesy	A1. Honesty A2. Altruism A3. Conciliatory Attitude A4. Modesty A5. Sensitivity with others
(C) Consciousness or Responsibility	C1. Competency C2. Organization C3. Sense of duty C4. The need to succeed C5. Self-discipline C6. Deliberation

RESEARCH VARIABLES

This section describes the details of the steps taken in the research. First, we obtained a set of data constituted by the psychological profile of the person and their musical preferences applying a statistical sample. Subsequently, the data was purified and the described data mining algorithms were applied to find relational patterns. Finally, the quality of the algorithms were evaluated and the results were analyzed. All steps mentioned above were framed within the SEMMA methodology. Below, the steps just outlined are explained with more details.

Phase 1 Sampling. The sampling was conducted for the inhabitants of the Metropolitan District of Quito (capital city of Ecuador) with the following characteristics: people aged between 15 and 35 years with mobile Internet access and smartphone.

Phase 2 Exploration. In this phase, variables for the research were defined. First the personality inventory (NEO-PI-R) was used in its version of 60-questions to measure the five major domains and six facets within each dimension.

The NEO PI-R profile indicates, on the upper side of the personal inventory, under the name of each scale, the raw score obtained for the scale in the discussion. The profile is generated on the basis of this gross score but visually displays T scores. The T scores are standardized scores, have a mean of 50 and a standard deviation of 10. A volume of 67% of all T scores are between 40 and 60. Theoretically, it is not possible to have T-scores below 20 or above 80, which is the reason why the visual profile does not show these intervals [5].

As shown in Fig. 2, the profile is delimited within five intensity categories for each score. The range of 20-35 indicates very low scores. The range of 35-45 indicates low scores. The range of 45-55 indicates mean scores. The 55-65 interval indicates higher scores. The range of 65-80 indicates very high scores.

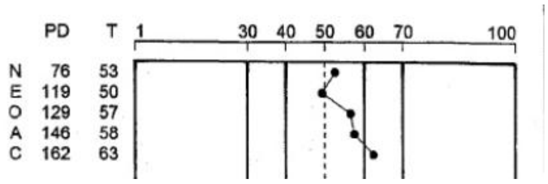


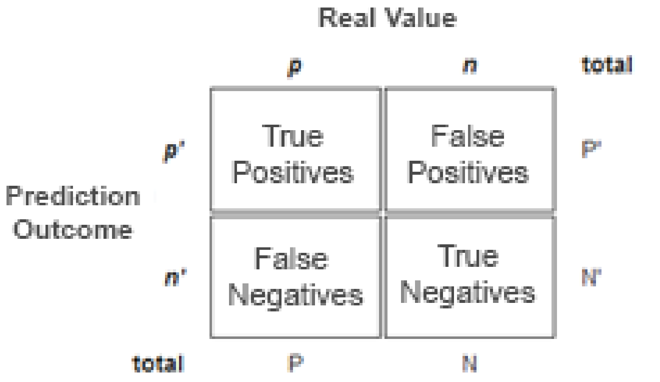
Figure 2: Scores to Define NEO-PI-R Case Profile

After obtaining the personality profile of the sample, the next step was to obtain the musical history and preferences of the person from Spotify. The 'getMyTop' method of Spotify's API was used to obtain the first twenty artists or tracks most heard by the user based on the 'Calculated Affinity'. Once the list of artists was obtained, the information was classified on the musical subgenre to which they belonged.

Phase 3 Handling. First, duplicate subgenre were eliminated, and then, through aggregation tasks, a new column was included with the generalization of the musical subgenres. Additionally, a column with a weight of 0 to 1 was added over the preference of certain genre. Finally, the database was partitioned into two sub sets: 80% for training of the classification algorithms and 20% for model performance tests.

Phase 4 Modeling. In this phase, an evaluation to the possible algorithms that are capable to manipulate dependent and qualitative variables were made. The analyzed algorithms for this research were Bayesian classifier, classification trees, random forest, and neural network (perceptron); Additionally, Clustering algorithm (K- Means) was applied to obtain profiles.

Phase 5 Assessment. In this phase, the performance of the different algorithms was evaluated according to the sensitivity and specificity of each one. After applying the contingency matrix, *p* values were obtained for each musical genre and a prediction *n* were generated by the model (see Fig. 3).



		Real Value		
		p	n	total
Prediction Outcome	p'	True Positives	False Positives	P'
	n'	False Negatives	True Negatives	N'
	total	P	N	

Figure 3: Possible values in the contingency matrix

The analysis of the ROC curve provided by the tools helped us to select the optimal models and discard the others. To draw a ROC curve, only the reasons for True Positive (TP) and false positive (FP) were necessary. Based on this, the performance values were obtained for each of the models (see Table 2).

Table 2: Performance of Data Mining Algorithms

Model	TP	FN	Precision	Error
Decision Tree	222	288	46.08%	53.92%
Naïve Bayes	190	320	35.69%	64.31%
Neural Network	200	310	40.98%	59.02%
Random Forest	228	282	45.69%	54.31%

To identify if there were similar traits in the groups, the K-means clustering algorithm was used in a range of 3 to 7 groups. Through this step, 4 groups with similar tendencies were obtained (see Figure 4). In addition, the dataset was associated with a weight field (value between 0-1) that indicated the percentage of preference of the person for a particular musical genre.

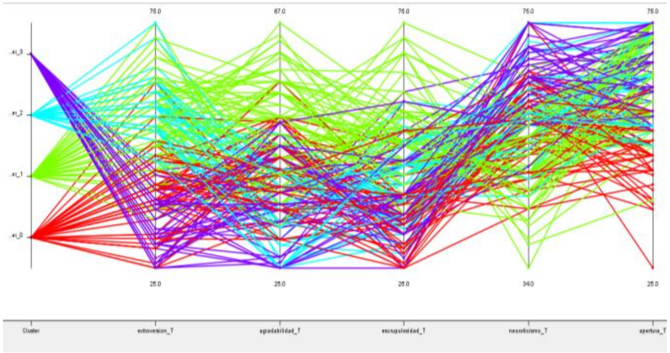


Figure 4: *Clusters and Personality Profiles*

EVALUATION OF RESULTS AND DISCUSSION

Based on the results obtained, it was possible to observe that the models with greater capacity of prediction are both the Decision Tree and the Random Forest. Clustering allowed differentiation of personality profiles for each cluster (see Table 3).

Table 3: Personality Profiles by Cluster

Factor/Cluster	0	1	2	3
Extroversion	Low	Medium-high	Medium-high	Low-medium
Affability	Medium	Medium-high	Low	Medium-low
Consciousness	Low	Medium-high	Medium	Low
Neuroticism	High	Medium-low	High	High
Openness	Medium	High	High	High

Through the analysis of the relationship between the clusters and the most representative musical genre of the sample, we can appreciate the following:

Rock: all clusters have a tendency to hear something of Rock (see figure 5).

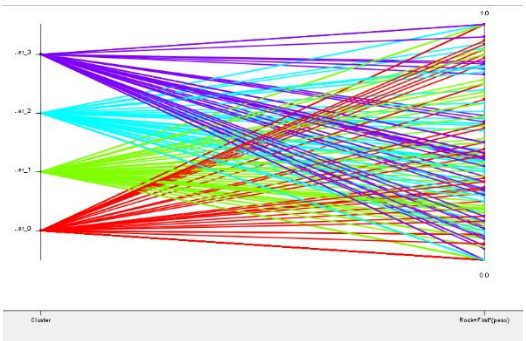


Figure 5: Relationship between clusters with genre Rock

Pop: All clusters listen to some pop music. Cluster 3 shows a certain tendency to dislike it (see figure 6).

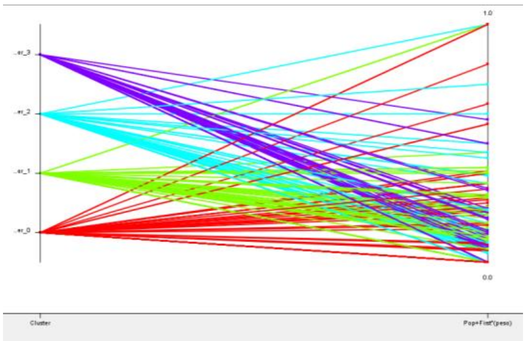


Figure 6: Relationship between clusters with genre Pop

Electronic: all listen to some electronics, cluster 1 has a greater preference for electronic music (see figure 7).

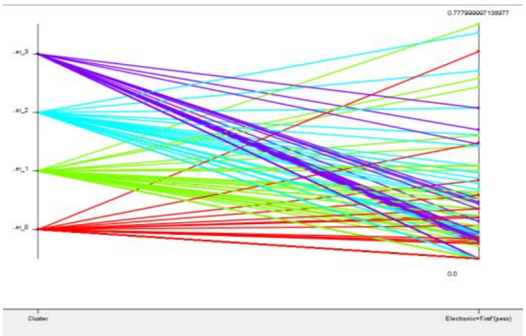


Fig. 7 – Relationship between clusters with genre Electronic

Latin: Cluster people 0, 1 and 2 like something of this kind of music, particularly those of cluster 2. Cluster 3 is not very affine to the latin musical genre (see figure 8).

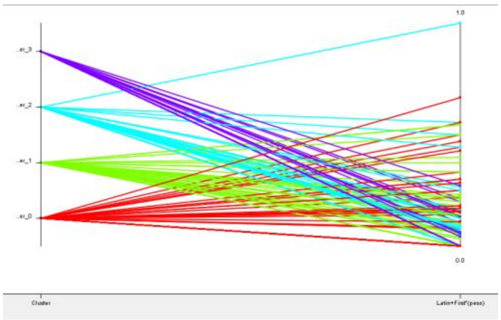


Figure 8: Relationship between clusters with genre Latin.

Dance: Cluster 1 and 2 have a predilection for the dance genre. Cluster 3 does not seem to like it much (Figure 9).

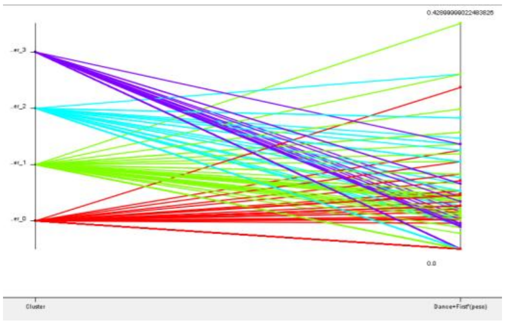


Figure 9: Relationship between clusters with genre Dance

Alternative: all clusters show some interest in this genre, particularly the clusters 1 and 3. (see figure 10).

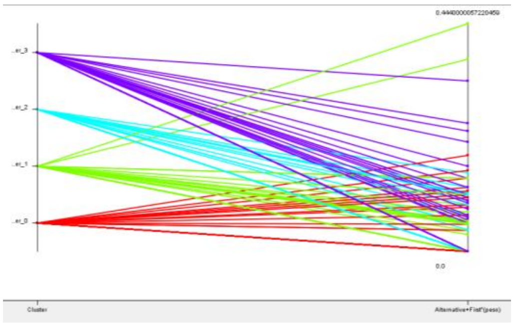


Figure 10: Relationship between clusters with genre Alternative

Having determined these possible relationships between personality and musical genre, we can conclude that the use of techniques and tools of data mining applied to musical databases allows to find certain associations between the musical preferences of a group of people and traits in their personality. However, these associations are not strongly determinant.

The task of determining or modeling a person's personality is very complex, since personality is a dynamic feature in humans and it is continually changing depending on their experiences and environment where they live. There are several inventories of personality with different benefits that try to explain this dynamic and have different benefits, but they are not conclusive and they are subject to a certain subjectivity. Changes in people's musical preferences is also very complex to determinate since these changes are subject to demographic, weather, and environmental issues.

Even though, we cannot determinate a strong conclusion, it was possible to establish relationships between the profiles and the musical preferences of the users which could be used in different fields such as marketing.

In the process of applying data mining tools, we found that the quality of input data is very important. In this aspect, we recommend to:

- Use sample of people from several cities and/or countries with pluricultural, pluriethnic, and different age's backgrounds.
- Delve deeper into the benefits of each personality inventory, and if possible, apply more extensive tests that could allow to delineate more deeply the personality of the individual and work with people who are willing, conscious and interested in the study.
- Search for different methods of musical classification. Although the musical genre is an alternative to try to classify music, it does not necessarily reflect the effect of music on the individual. There are other mechanisms to classify it such as some methods assisted by machine learning, digital signal processing, classification by emotions, and so on, which can shed other relationships between personality and music.

CONCLUSIONS

The influence of music on the behavior and psycho-emotional construction of people during their life at the personal, social, and cultural levels is undeniable. This paper has identified a mechanism capable of evaluating personality traits through five major factors centered on the OCEAN model and different techniques of data mining to find the associations between musical preferences and personality traits.

REFERENCES

[1] Asphalt. 2013, "Music in Human Evolution, " Retrieved: Feb. 16, 2016 from Melting Asphalt: http://www.meltingasphalt.com/music-in-human-evolution/

[2] Koelsch, S., 2016, "Brain and Music: A Contribution to the Investigation of Central Auditory Processing with a new Electrophysiological Approach," Retrieved: Feb. 16, 2016 from http://pubman.mpdl.mpg.de/pubman /item/escidoc:720506/component/escidoc:720505/koelsch.pdf

[3] Goldberg, L. R., 1993, "The structure of phenotypic personality traits," American Psychologist, 48, pp. 26-34.

[4] Goldberg, L. R., 1990, "An alternative "description of personality": the big-five factor structure," Journal of Personality and Social Psychology, Vol 59(6), pp. 1216-1229.

[5] Costa, McCrae, 1992, "Revised NEO Personality Inventory (NEO-PI-R) and NEO Five Factor Inventory (NEO-FFI) manual," Odessa, FL: Psychological Assessment Resources.

[6] Martinez, A. I., 2002, "Temperamento, carácter, personalidad. Una aproximación a su concepto e interacción, " Revista Complutense de Educación, Vol. 13 No. 2 , pp. 617-643.

[7] Samuel D., Gosling, P., 2003, "A very brief measure of the Big-Five," Journal of Research in Personality, 37, pp- 504–528.

[8] Ramageri, B. M., 2010, "Data Mining Techniques and Applications," Indian Journal of Computer Science and Engineering, Vol. 1, No. 4, pp. 301-305.

[9] Pearson, J. L., Dollinger, J. S., 2004, "Music preference correlates of Jungian types," Personality and Individual Differences, Volume 36, Issue 5, pp. 1005-1008.

[10] Rawlings, D., Ciancarelli, V., 1997, "Music Preference and the Five-Factor Model of the NEO Personality Inventory," Psychology of Music. Volume: 25, pp. 120-132.

[11] Delsing, M. J., Bogt, T. F., Engels, R. C., Meeus, W. H., 2008, "Adolescents' music preferences and personality characteristics," European Journal of Personality, Vol. 22, Issue 2, pp. 109–130 .

[12] Schwartz, K. D., Fouts, G. T., 2003, "Music Preferences, Personality Style, and Developmental Issues of Adolescents," Journal of Youth and Adolescence, Vol. 32, Issue 3, pp.205–213.

[13] Rentfrow, P. J., Gosling, S. D., 2003, "The Do Re Mi's of Everyday Life: The Structure and Personality," Journal of Personality and Social Psychology, Vol. 84, No. 6, pp. 1236 –1256.

[14] Ferrando, M. G., 1986, "El análisis de la realidad social, Métodos y técnicas de investigación," Madrid: Alianza.

[15] SAS, 2008, "SAS Enterprise Minner: SEMMA," Retrieved from http://www.sas.com/content/dam/ SAS/en_us/doc/factsheet/sas-enterprise-miner-101369.pdf

[16] Aiken, L. R., 2003, "Test psicológicos y evaluación," 11th edition, México: Pearson, Prentice Hall.