ISSN 1870-4069

Building a Dataset for Professional Interest Recognition

María Lucia Barrón-Estrada, Ramón Zatarain-Cabada,
Abel Robles-Montoya, Héctor Manuel Cárdenas-López,
Arcelia Judith Bustillos-Martínez, Néstor Leyva-López

Tecnológico Nacional de México,
Instituto Tecnológico de Culiacán,
Mexico

{lucia.be, ramon.zc, arcelia.bm}@culiacan.tecnm.mx

Abstract. The personality of the student is a key factor influencing the choice of career as students choose their career according to their personality. However, young people do not always know which options are the most suitable for their own characteristics. This paper presents the design and implementation of an intelligent system to help the student to better choose the professional career to study using a video of the same student as input. For the development of this system, a dataset was used that was created with personality and professional interest questionnaires, videos, and information related to the academic and professional training of the participants. This dataset was used to train machine and deep learning models that can predict personality and career interests from video. The results of the personality prediction models were in the order of 0.18, while for the model of professional interests it was 0.16.

Keywords: Automatic personality recognition, professional interests, RIASEC, HEXACO, machine learning.

1 Introduction

The term "dropout" is used to describe students who stop attending school for a variety of reasons, including personal obligations, academic failure, or the expiration of their enrollment [1]. In Mexico, however, the Secretaría de Educación Public (SEP) uses the term to describe students who stop attending school without considering their academic performance, choice of careers, or other factors.

We refer to school dropouts throughout this essay using the first term. Only 6 out of 10 Mexican students enrolled in higher education (university and technology bachelor's degree) can complete their studies, according to data released by SEP (2019-2020) [2]. The socioeconomic and personal factors are the most frequently cited causes for the 40% of students who drop out of university.

However, other significant factors include academic dissatisfaction, poor school performance, and career choices [3]. According to Holland's theory, students choose academic environments in accordance with their personalities [4], but young people don't always know which options are suitable for them considering their own interests.

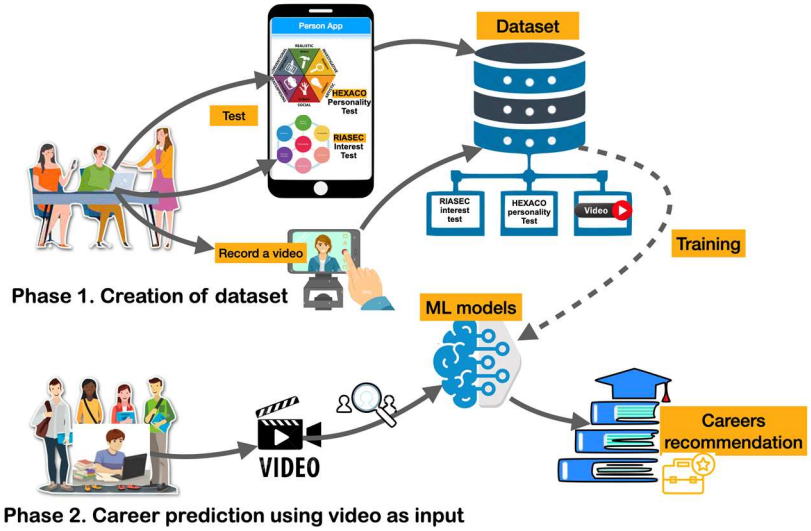

Fig. 1. General system to predict career based on personality and interest.

The personality of the student has been identified as a factor that influences its career choice. Therefore, career counseling using instruments (such psychometric tests) can help students discover a career appropriate to their personality and therefore prevent school dropout.

This can lead to a number of circumstances, such as not obtaining an adequate academic performance in that career or losing interest. Numerous academic fields, including psychology, philosophy, and others, have extensively examined personality analysis [5]. The computer vision community has recently showed interest in identifying personalities, starting with physical information expressed through the face, postures, or behaviors of an individual [6].

The personality analysis can benefit society in a variety of other ways as well. For instance, it can be used in education to determine the best method of instruction; in business when hiring employees, it can be used to choose candidates who have the right personality for the job; and in educational guidance, it may be used to assist young people in choosing their field of study to reduce the number of school dropouts.

Vocational guidance is a collection of procedures focused on the vocational dilemma that tries to give each subject the resources necessary to make the best decision possible given their circumstances [7]. To help people discover more about themselves and choose a job they enjoy while also achieving a gratifying performance, vocational counselling is now backed by approaches and tools like interviews, surveys, or psychometric testing [8].

The Kuder Vocational Interest Scale, the Explora Test, and the Basic Academic and Professional Interest Areas Questionnaire are some of the vocational instruments suggested by various authors; each one uses a different methodology to identify the area or profession that best suits each person [9]. A software system was created that, through personality recognition using videos, can predict which professional fields of study are most suitable for each student.

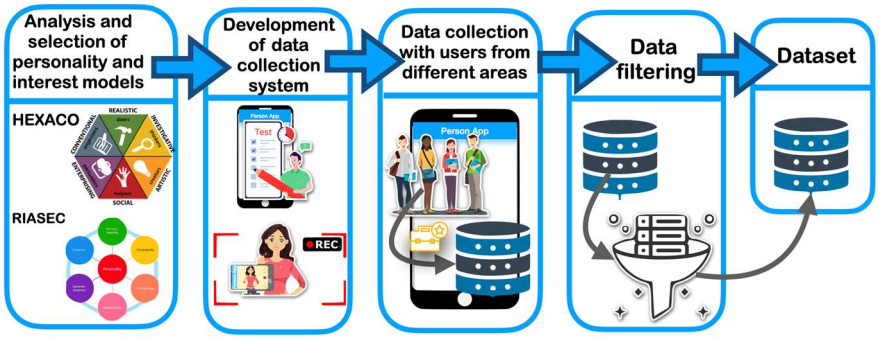

Fig. 2. Methodology for the creation of the personality-interest corpus.

This is because school dropout is a problem, and because academic dissatisfaction and poor school performance are linked to a poor career choice. Two artificial intelligence (ML) models are required for this process to make predictions.

The first model was trained using user-recorded videos and personality questionnaires in such a way that the model received the videos as input and generated as output the user's personality attributes; the second model received the personality attributes and responded with the user's vocational interests.

By combining the two models, it is possible to predict the user's vocational interests based on the personality attributes. The key obstacle to attaining it is the need for a large dataset of user videos, personality tests, and career interests. This dataset needs to have enough records and high-quality data to enable the models to train.

In this paper, a mechanism for predicting a user's associated vocations by personality detection in videos is presented. The paper structure is as follows: section 2 shows related works that were analyzed in this research; in section 3 we explain the method to develop this research work; section 4 presents the results that were obtained; and finally, section 5 shows conclusions and limitations.

2 Related Work

Interest has been the subject of studies in vocational psychology that have been applied in actual situations, including the First World War. Academics had known this for most of the previous century, and American industry swiftly embraced its significance in choosing employment. The first instrument to measure people's interests was created in 1919 by Clarence S. Yoakum at Carnegie Institute of Technology. He used a logical sampling strategy meant to represent the domain of interests [4, 10].

Holland's theory, which states that it is possible to categorize each person in one of six different dimensions (RIASEC) —Realistic, Investigative, Artistic, Social, Enterprising and Conventional—or a combination of these—has been one of the most influential for the development of various questionnaires to assess interests [4, 11]. This hypothesis has inspired the creation of numerous tools throughout the years, including questionnaires that assess each person's interests.

Table 1. PersonApp 2.0 Functional requirements.

Requirement	Description
RF1	The system will allow users to take a standardized vocational test.
RF2	The system will allow users to take a standardized personality test.
RF3	The system shall store user information.
RF4	The system will show the user a description of his or her vocational interests.
RF5	The system will show the user a list of careers according to their interests.
RF5	The system will show the user a list of careers according to their interests.
RF6	The system will allow downloading a report of the results in PDF format.

The Department of Labor in the United States created O*NET, a career exploration and planning tool, in 1999 using Holland's six kinds as a benchmark. This platform provides a condensed version of the interest questionnaire that only accounts for 10 questions for each RIASEC category and totals 60 questions [12, 13].

Hansen [13] then examined the psychometric properties of a condensed form of the interest questionnaire and discovered evidence of its reliability, validity, and excellent concordance with the full version.

The U.S. Department of Labor uses this survey under a free-use license.There have been several, long running, and diverse attempts to identify a model that might accurately capture human personality on a global scale [14]. The Big Five model, which consists of five dimensions that each measure a different personality attribute, is one of the most extensively used models to describe how people differ from one another. The five dimensions—Openness, Consciousness, Extraversion, Agreeableness, and Neuroticism—are collectively referred to as the OCEAN model.

Each dimension is symbolized by its first letter. K. Lee and M. Ashton created the HEXACO-PI-R instrument, which Roncero Fornés and Belloch translated into Spanish with the authors' approval [14]. The validity of the Spanish version of this instrument was examined, and it was determined that the internal consistency and stability were very satisfying.

They came to the conclusion that HEXACO is a suitable tool to evaluate personality. Since the authors claim that this model adds a new dimension called Honesty that completes all the features of a person and increases interpretability, it has lately been researched for its theoretical advantages over the most widely used model of the big five, OCEAN [15].

The six dimensions of the HEXACO model are as follows: Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness and Openness to Experience [14, 16]. In two separate studies, verko & Babarovi [17] describe the relationship between personality and interests.

In the first study, 602 high school graduates completed the HEXACO model using the HEXACO-PI-(R)-100 questionnaire and the OCEAN model using the IPIP-50 questionnaire, and in the second study, 981 participants completed the PGI-Short and HEXACO-60 questionnaires.

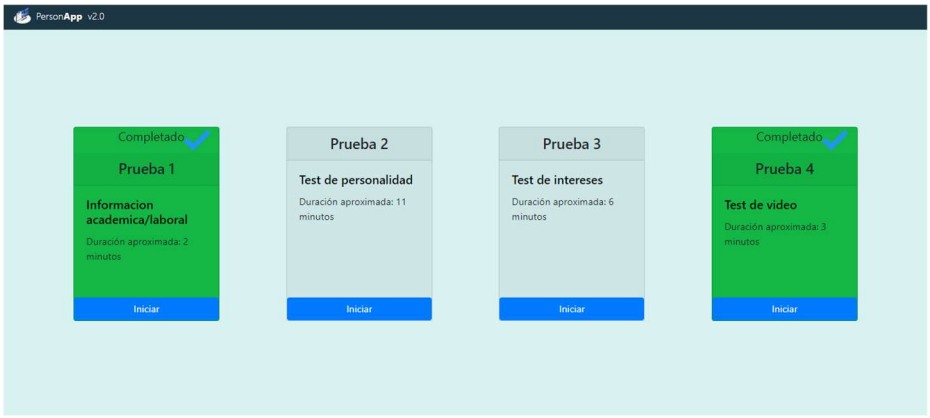

Fig. 3. PersonApp main menu.

Positive or negative correlations between interests and personality were found in the results of both investigations; also, the HEXACO model shown predictive advantages in describing interests compared to the Big Five.

In a recent study, Barmanpek [18] discovered that personality and interests are generalizable and do not vary across time, groups, or cultures. He also discovered some direct relationships on some RIASEC and HEXACO model variables. Tandera, et al. [19] describe a technique to estimate personality using the user's Facebook information and the OCEAN model.

In this study, they evaluate the outcomes of using machine learning and deep learning and concluded that the latter is superior to the former since deep learning can increase the accuracy of the majority of the features.

By categorizing participants of a video social network and then utilizing the Big Five model to identify, frame by frame, the emotion present in each facial expression, Biel, et al. [20] take advantage of state-of-the-art technologies for facial expression identification.

They concluded that these can be helpful for predicting personality after discovering that particular facial expressions are closely associated to some personality traits. Using text mining algorithms, Pavithran and Ashraf [21] try to determine career affinity. The theoretical foundation for this research, which makes use of the O*NET career database, is Holland's RIASEC model.

The method entails removing text from Twitter, preprocessing the words, labeling them in an RIASEC dimension, counting the words, and so on. Walker, et al. [22] present the "Basel Face Database", which integrates three various image gathering methodologies for personality recognition. To display variations in the dimensions of the big two (communion and agency) and the big five, images of real faces were used to create this database.

The procedure involved selecting participants from Amazon Mechanical Turk and obtaining their agreement for data collection before displaying a random selection of both original and changed photographs and asking them to evaluate the images by ranking them according to each personality dimension.

3 Methodology

A new approach was adopted to carry out this project that made use of an intelligent system. Like a standardized vocational orientation test, it uses video analysis to identify the user's personality and traits before recommending a career. The benefit of this technique is that students simply need to film a little video instead of filling out a lengthy questionnaire.

Fig. 1 provides a general overview of the system and clearly illustrates the system's two phases: phase 1 involves the creation of a dataset for training machine learning models, and phase 2 involves making career recommendations using user videos as input. The HEXACO model's and the RIASEC interest model's personality traits must be used in the system's implementation.

A platform is utilized to gather data from many users by distributing personality and interest questionnaires and capturing videos through a few trials, as shown in Fig. 1, phase 1. A dataset will be created from the data collected using that platform and used to train neural networks. One of them will analyze video to predict personality using questionnaires, while the other will use personality to predict career preferences.

This method is suggested rather than predicting interests directly from video due to the vast research that already exists on automatic personality detection. By combining these models, it is possible to predict career interests from videos (see Fig. 1, phase 2).

A four-stage methodology was created to develop a dataset for personality and interest recognition based on the research examined on the generation of corpus or datasets for attribute prediction, as illustrated in Fig. 2. The following sections detail each stage of the methodology.

a) Analysis and Selection of Personality and Interest Models.

The background information on personality and interest models was examined to determine which model would be most useful for the proposed situation. We chose the most well-known and extensively researched psychological models. Finally, it was confirmed that there were questionnaires with research aims for each model since the goal is to predict professional interests using personality. Both models needed to have direct links between their traits or the dimensions they cover.

In light of the aforementioned requirements and the background data, it was discovered that Holland's theory can accurately represent interests and also has connections with several aspects of other personality models. In addition, the O*NET database uses the RIASEC model that Holland proposed, and for these reasons, this model was chosen to assess professional interests.

The HEXACO model was chosen to assess personality because to advancements over the conventional Big Five model as discussed in the background and, more significantly, due to evidence of a greater association between the traits represented in HEXACO and the RIASEC interest model.

b) Development of the Data Collection System (videos and questionnaires).

With the use of the OCEAN model, a software program called PersonApp [23] was created to automate personality questionnaires and collect data while also producing a video dataset.

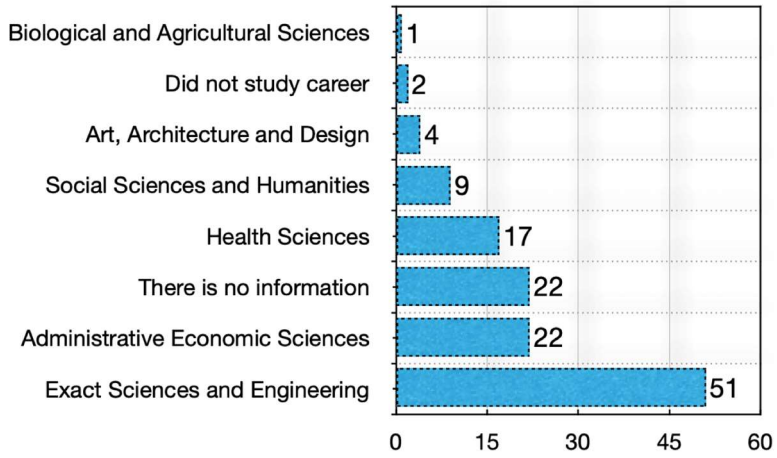

Fig. 4. Number of records by area.

The system specifies three experiments for the filming of the videos, each of which requires the user to discuss a different subject for the purpose of four quick 15-second recordings. The IPIP-50 is the used personality test. A client-server architecture was used to construct this system utilizing a layered architectural paradigm.

The PersonApp program served as the foundation for the creation of a new version that considers the updated specifications created to produce the datasets of videos, personalities, and professional interests. The updated version, known as PersonApp 2.0, integrates the functional specifications listed in Table 1 that were discovered during the analysis and model selection stage.

The following modifications were considered for PersonApp 2.0. As shown in Fig. 3, a monitoring interface was first added so that the user could see his/her progress as he/she finished each test.

The RIASEC interest test, which consists of 60 multiple-choice questions on users' interests and 2 control questions that direct users to choose a certain response to ensure that the results aren't random, was also added. For the responses, we utilized a Likert scale with five possibilities, where the first one has a value of 1 and goes up to 5 for the last option (1: I like it very much, 2: I dislike it, 3: I am not sure, 4: I like it, and 5: I like it very much). Each question in the questionnaire has five possible answers.

The third modification to the personality test replaced the IPIP test that was previously used to determine the users' OCEAN model values. This test was replaced by HEXACO, which has 72 questions with 5 possible answers (1: Totally disagree, 2: Partially disagree, 3: Neither agree nor disagree, 4: Partially agree, 5: Totally agree), of which 70 are about the test and 2 are control questions.

The fourth update was the addition of many questions to gather information about some of the participants' academic backgrounds, including their schooling and professions, positions held, and others. The gathering of data was performed in the non-relational Google Firebase database. The values that correspond to each model are determined when the user completes each test, and these values are then translated into a range from 0 to 1 to be saved in the database.

Table 2. Number of records by gender.

Gender	Number of records
Female	36
Male	70
Prefer not to say	2
No information	20

Additionally, information regarding the user's response time, how many of the control questions they successfully answered, user reviews of the program, and videos of the three trials are all preserved. Tests with several individuals were conducted before the application for database generation was released to ensure proper performance.

c) Data Collection with Users from Different Areas.

A group of 30 experts from various fields were given the app to test how well it worked across various platforms. To fill out the tests, certain individuals from various backgrounds (mainly students and professionals) were chosen, resulting in a sizable percentage of individuals with comparable personality traits in the database. For the prediction of interests to be possible, this is crucial.

Participants were chosen from among classmates, friends, and family members for the initial stage of data collection, which started with gathering as much information as possible without paying close attention to maintaining a balanced number of entries for each subject area.

To adopt a procedure centered on data collecting and expand the number of entries, undergraduate students from the Instituto Tecnológico de Culiacán joined the project in a later stage. For this iteration, participants were sought from fields with less items in the database.

d) Data Filtering.

To ensure that it can be used to train artificial neural networks, all data gathered into the dataset must pass a filtering process. They must accurately reflect the professional interests and personality of the users; otherwise, no neural network or other type of technique will be able to find the relationship between these variables. It is crucial to remove the inconsistent data that indicate random replies for the dataset to satisfy the needs of this job.

Two mechanisms were used to identify these cases:

- Categorize the amount of time the user spends responding to the questionnaire's questions. Outliers were eliminated, and a normal curve was plotted to determine that, in the personality and interest tests, 210 and 120 seconds, respectively, are the minimal times at which it is assumed that the user completed the questions diligently.
- Verify control question responses. The user is given instructions to select a specific response, such as "I really dislike it," and if they select any other response, it is assumed that they answered the questionnaire at random.

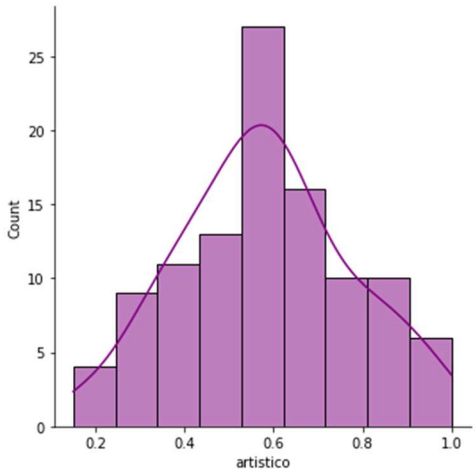

Fig. 5. Graph of the distribution of aristic attibute.

4 Results

The launch of the app gave participants a chance to voice any feedback or issues they had with the setup. The following major adjustments were suggested by users who gave us feedback on the platform, we also found areas for improvement, as follows:

- A warning note was added at the start to inform the user that he/she must agree to participate and that their information would only be used for research purposes.
- Users of the app can view their progress on a menu.
- The software has a feature that allows you to pause video recording and restart it.
- Visual components were added to help the user know when an experiment's recording is complete.
- An additional counter was included to show the recording time.
- Additionally, email validation was added, which the user must confirm in order to access their account.
- Also included is a screen that instructs consumers on which option to select based on their circumstances.

The data on the users' academic backgrounds was used to create the graph in Fig. 4, where it is possible to see the number of records for each area of knowledge. Given that 50% of the data came from the sciences and engineering, these fields attracted most participants. In contrast, only a small number of people engaged in the fields of biological and agricultural sciences, as well as art, architecture, and design.

It will be crucial to attempt and collect additional data from the areas with less records because the uneven data could make it challenging to develop a model that can generalize effectively. Table 2 lists the number of records broken down by the gender of those who responded to the app's questionnaires.

Table 3. Descriptive statistics for each attribute in HEXACO and RIASEC.

Attribute	Mean	St. dev.	Min	Max
Neuroticism	0.44	0.11	0.16	0.7
Honesty-Humility	0.63	0.07	0.48	0.76
Emotionality	0.67	0.09	0.46	0.9
Extroversion	0.68	0.09	0.42	0.9
Agreeableness	0.68	0.08	0.52	0.88
Conscientiousness	0.65	0.07	0.44	0.76
Openness to experience	0.68	0.08	0.48	0.84
Realistic (doers)	0.51	0.18	0	1
Investigative (thinkers)	0.61	0.2	0.2	1
Artistic (creators)	0.58	0.19	0.15	1
Social (helpers)	0.6	0.18	0.1	0.97
Enterprising (persuaders)	0.58	0.2	0	1
Conventional (organizers)	0.54	0.2	0	1

Like the knowledge domains, there is a significant disparity in the data; in this case, there are approximately twice as many men as women. Figure 5 is a graph that represents the results of the personality test's Artistic attribute, all other attributes follow a similar pattern. These graphs demonstrate how the data fit into a normal distribution quite well, however the ranges for the personality tests are not comprehensive.

This can have a severe impact on how these data are used to train machine learning models because, in the specific case of neuroticism, it has a range from 0.16 to 0.7. This pattern is reproduced in the other traits of the HEXACO model, where there are also incomplete ranges, but in different measures.

The PersonApp 2.0 system created 215 accounts, of whom 193 gave information on their education and employment, 174 completed personality and interest questionnaires, and 98 participated in video experimentation. The appropriate filters were applied to each questionnaire, yielding 127 records for the final database.

Some data about the dataset are included in Table 3. The first column includes all the HEXACO and RIASEC model features, as well as neuroticism, which is an OCEAN model attribute. The table only considers filtered data from 128 participants, with maximum and lowest values of 1 and 0 for each attribute, respectively.

5 Conclusions

In this study, the issue of school dropout was discussed, along with how computing may help to enhance current solutions. The work in this paper, presents the creation of a platform where users can take personality and career interest tests and produce videos.

Using this platform, information was gathered by using specific filters to eliminate any incomplete or incorrect information, and ultimately, a database was created that had 127 questionnaires and 98 videos.

When conducting these types of experiments, using a digital platform can be much more helpful than using other, more rudimentary methods (such as filling out questionnaires and filming videos in person) because the data can be collected, stored, and organized automatically. However, it would be important to have more mechanisms in place to ensure the accuracy of the data.

The dataset has some drawbacks, including an uneven distribution of people by sex and by various types of expertise. As a result, the number of records is not equal for each area and for each gender. Another issue resulting from the unbalanced data is that since most people have similar traits, such as gender and career, the values for each attribute fall within a narrower range.

Since the data are unbalanced, there are no records for the whole range of the data, even if each characteristic measured has values ranging from 0 to 1. Because they require balanced data to explore for links between variables, rather than just one or two, this issue makes prediction models ineffective.

Future work will continue to expand the database to achieve a better balance of the experiment participants' characteristics. Additionally, the project's next phases will see the development and optimization of prediction models for both personality and professional interests in order to use them in a system that can take user videos as input and recommend products to them.

References

1. Aldaco-Linares, R. N., Carpio-Hernández, M. M.: Las estrategias para abordar el abandono escolar en una institución de educación superior tecnológica en México. In: Congresos CLABES (2017)
2. DGPPYEE-SEP. (s/f). Gob.Mx. Recuperado el 14 de marzo de 2022, de http://www.planeacion.sep.gob.mx/estadisticaeindicadores.aspx.
3. Reyes, J.: Modelo de decisión multicriterio para la selección de carrera universitaria. Investigación y Desarrollo, vol. 6, no. 1, pp. 25–32 (2013)
4. Hansen, J. C.: Interest inventories. Handbook of psychological assessment, pp. 169–190 (2019) doi: 10.1016/B978-0-12-802203-0.00006-7
5. Sinisterra, M. M., Cruz, J. P., Gantiva, C.: Teorías de la personalidad. Un análisis histórico del concepto y su medición. Psychologia. Avances de la disciplina, vol. 3, no. 2, pp. 81–107 (2009)
6. Jacques-Junior, J. C. J., Güçlütürk, Y., Pérez, M., Güçlü, U., Andujar, C., Baró, X., Jair-Escalante, H., Guyon, I., van-Gerven, M. A. J., van-Lier, R., Escalera, S.: First impressions: A survey on vision-based apparent personality trait analysis. IEEE Transactions on Affective Computing, vol. 13, no 1, pp. 75–95 (2019) doi: 10.1109/ TAFFC.2019.2930058
7. Reyes-Campos, I. M., Novoa-Cely, A. M.: Orientación vocacional. Universidad Central (2015)
8. Ramos-Monsivais, C. L., González, B. A.: Orientación vocacional, aprendizaje socio-emocional y sentido de vida en la educación superior. Dilemas contemporáneos: educación, política y valores, vol. 8, no. SPE5 (2020) doi: 10.46377/dilemas.v8i.2500

9. Alfaro-Barquero, A., Chinchilla-Brenes, S.: Diseño de un instrumento de preferencias vocacionales en administración, materiales y biotecnología. Revista Costarricense de Psicología, vol. 38, no. 2, pp. 99–124 (2019) doi: 10.22544/rcps.v38i02.01
10. Holland, J. L.: Making vocational choices: A theory of vocational personalities and work environments. Psychological Assessment Resources (1997)
11. Acosta-Amaya, M. M. La orientación profesional: Revisión del modelo RIASEC y la teoría social cognitiva del desarrollo de carrera. 2 cuadernos de Ciencias Sociales: Investigación en Psicología, vol. 47 (2018)
12. Rounds, J., Su, R., Lewis, P., Rivkin, D.: O* NET interest profiler short form psychometric characteristics: Summary. Raleigh, NC: National Center for O* NET Development, pp. 1–43 (2010)
13. Hansen, J. I. C.: Interest inventories. Handbook of psychological assessment, pp. 169–190 (2019) doi: 10.1016/B978-0-12-802203-0.00006-7
14. Roncero, M., Fornés, G., Belloch, A.: Hexaco: Una nueva aproximación a la evaluación de la personalidad en español. Revista argentina de clínica psicológica, vol. 22, no. 3, pp. 205–217 (2013)
15. Ashton, M. C., Lee, K.: Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. Personality and social psychology review, vol. 11, no.2, pp. 150–166 (2007) doi: 10.1177/1088868306294907
16. Lee, K., Ashton, M. C.: Psychometric properties of the HEXACO personality inventory. Multivariate behavioral research, vol. 39, no. 2, pp. 329–358 (2004) doi: 10.1207/s15327906mbr3902_8
17. Šverko, I., Babarović, T.: Integrating personality and career adaptability into vocational interest space. Journal of Vocational Behavior, vol. 94, pp. 89–103 (2016)
18. Barmanpek, U.: Relationship between personality traits and vocational interests: A multi-faceted study. Doctoral dissertation, University of Leicester (2019)
19. Tandera, T., Suhartono, D., Wongso, R., Prasetio, Y. L.: Personality prediction system from facebook users. Procedia computer science, vol. 116, pp. 604–611 (2017) doi: 10.1016/j.procs.2017.10.016
20. Biel, J. I., Teijeiro-Mosquera, L., Gatica-Perez, D.: Facetube: predicting personality from facial expressions of emotion in online conversational video. In: Proceedings of the 14th ACM international conference on Multimodal interaction, pp. 53–56 (2012)
21. Walker, M., Schönborn, S., Greifeneder, R., Vetter, T.: The basel face database: A validated set of photographs reflecting systematic differences in big two and big five personality dimensions. PloS one, vol. 13, no. 3, pp. e0193190 (2018)
22. Bátiz Beltrán, V. M. (2021). Sistema de evaluación de la personalidad y de las emociones en el proceso cognitivo. Tesis de Maestría en Ciencias de la Computación. Instituto Tecnológico de Culiacán.