





Risk Profiles of Financial Service Portfolio for Women Segment Using Machine Learning Algorithms

Jessica Ivonne Lozano-Medina, Laura Hervert-Escobar^(✉) ,
and Neil Hernandez-Gress 

Tecnologico de Monterrey, 64849 Monterrey, NL, Mexico
laura.hervert@tec.mx
<http://tec.mx/en/>

Abstract. Typically, women are scored with a lower financial risk than men. However, the understanding of variables and indicators that lead to such results, are not fully understood. Furthermore, the stochastic nature of the data makes it difficult to generate a suitable profile to offer an adequate financial portfolio to the women segment. As the amount, variety, and speed of data increases, so too does the uncertainty inherent within, leading to a lack of confidence in the results. In this research, machine learning techniques are used for data analysis. In this way, faster, more accurate results are obtained than in traditional models (such as statistical models or linear programming) in addition to their scalability.

Keywords: PCA (Principal Component Analysis) · Portfolio optimization · Machine learning

1 Introduction

In Latin America and the Caribbean (LAC), approximately one in four are poor and one in ten cannot meet their basic food needs. The latter being especially true for children. According to The World Bank [2], even the most equal country in LAC (Uruguay) is more unequal than the most unequal country of OECD (Portugal). But it has been found that women have been important in the reduction of poverty across LAC countries. According to the study [2], it was found that women made crucial contributions to extreme and moderate poverty reduction between 2000 to 2010. Growth in female income translated to 30% reduction in extreme poverty along with 39% reduction by male income. Overall, female labor market income was twice as effective in reducing the severity of poverty compared to male labor market income.

Some benefits of the inclusion of women include:

- More resiliency against poverty and better coping mechanisms during economic shocks, especially in cases of dual-income households.
- Increases in female labor income and labor market participation seem proportional to higher enrollment rates and a higher education (closer to men).

Some of those studies, have found that women tend to be more risk-averse than men in their financial management [6]. Though there is much controversy if the difference is significant or invalid [5], the issue of understanding the financial attitudes of people is of interest for many institutions. Such institutions include governmental organizations, who wish to increase the involvement of women in the finance industry [23], and also of financial institutions that wish to classify better the profiles of their clients.

Currently, banks manage their finance portfolio clients depending on a risk profile. These profiles are unique for every client. What this means is that every client has a different risk tolerance. That is the level at which they are still comfortable risking money, and it depends on the lifestyle, age and personality of the person. Classifying these profiles correctly helps the client have decent earnings while having peace of mind [15].

In the past, many studies have been done to determine if there are differences in risk management between genders. For example, there has been numerous studies to prove or disprove such claim [6]. This included data gathering from experiments meant to measure that difference, and from experiments that were not meant to measure it. Yet these studies prove time and again that there is truth to the claim.

On the other hand, there has been previous attempts to model, the classification of woman financial habits. Studies include analysis on household income census from Italy [3], Sweden [12] and United States [21]. The majority of these studies use linear programming techniques [12,17,26] and probit models [3,18].

One of the studies has used an unsupervised machine learning technique. Simms [21] used Cluster Analysis to analyze women habits on financial advice and was able to identify two prominent groups. Cluster Analysis is useful for unlabeled data, that is instances that don't have a class assigned. And these type of techniques can benefit the study of risk profiles due to being able to create complex models, manage the amount of features and classify them with high accuracy.

Therefore as a means to understand the segment better, this research seeks to generate an adequate profile of women segment and explain the variables and indicators of the profile. If this model proves to be efficient identifying risk and classifying, it will provide a better understanding of the client risk profile of the target segment and a better-tailored product. This will help develop entrepreneurial woman in Mexico that seek to grow their financial assets and generate a reasonable income from their investments. For the financial institution it will help to provide a better product, a safer investment, and aid in its decision-making as a better way of classifying and understanding user risk profiles. The rest of the document is organized as follows: Problem definition and context are given in Sect. 2. Section 3 presents methods and approaches used. Finally Sect. 4 presents the conclusions and future work.

2 Problem Definition in Mexican Population Context

Modelling human behavior is a complicated task. Some examples are the modelling market price movements [9], manage credit card customers [10], and forecasting financial risk [24]. Most of them are done with statistical and operational techniques [22].

In finance, the modeling of the behavior of potential clients and their risk, in general, does not distinguish gender. However, analyzing the women's segment separately represents an opportunity for financial companies to expand their portfolio and contribute to gender equality.

Women are among the most vulnerable groups and provide much concern for governments in the world. According to the UN [23], at least 50% of the women in the world have a paid wage and salary employment, which is an increase of 40% compared to the 1990s. Yet women earn 24% less than men when having the same work. Therefore there is still much work to do to provide gender equality, poverty eradication and inclusive economic growth. Women make enormous contributions to economy either by working in a business, on a farm, as entrepreneurs or employees, or by doing unpaid care work at home, and still there is much to do to aid women in their development.

2.1 Mexican Financial Habits of the Population

In the context of Mexico, The National Institute of Statistics and Geography (INEGI) released the results of a survey that reveals the differences between being men or women in Mexico [11]. The survey contains information on population between 18 and 70 years old. The analysis was made for rural and urban areas and includes some topics related with habits in the expenses management. Some examples are: having a savings account, having an expense record book, source of your income (foreign or national), etc.

Saving Accounts. Mexican women overall use less financial instruments such as retirement savings accounts. Being the participation of 20% in urban areas and 15.2% in rural areas. This result is due to they have less presence in the workforce. Typically, Mexicans contribute to their retirement account through contributions from their salary. This contribution is made automatically (by law) to each salaried worker. People who do not have an employer should look for ways to save for retirement. Women overall did not have a saving account due to not having a job (43.8%) or not knowing about the benefit (18.9%).

Type of Savings. Three options were considered: formal savings (within a financial institution), informal savings and a combination of both. Overall, 6 out of 10 Mexicans have informal savings. Also, 2 out of 10 do not have savings and about the same amount have formal savings. In urban areas, men combined informal and formal savings, while women prefer informal savings. Contrary so, in rural areas, men prefer informal savings while women choose formal savings,

or a combination of both. Additionally, in rural areas, women have more access to formal credits than men, while in urban areas it is the opposite.

National Business Trends. Businesses in Mexico can be divided in four categories: Commerce (46.7%), services (39.1%), manufacturing (12.2%) and other types of economic activities (2%).

While 71.8% of the workforce can be found in the areas of commerce and services in 2019, the best paid jobs can be found in the areas of manufacturing with an average of 8,389.9 dollars per year per worker. Men have the largest presence in the areas of manufacturing, where 63.6% of the men in the workforce can be found. While women have the largest presence in the areas of commerce, with 46.9%. The states with the highest participation of women are the ones in south-center of the country, specifically in Guerrero (49.8%) and Oaxaca (48.6%). Some problems facing companies in Mexico are the insecurity (39.1%), very high costs (23.9%), and the unfair competition 21.1%.

Mexico has a dominant type of business, named micro-businesses. They are the most common type found in all the country. They give employment to more than a third of the citizens. Usually, the workforce is maximum 10 people or less. They represent 95% of all businesses, contribute 37% of the total employed workforce and 14.2% of the incomes.

From census, it has been found that most businesses are located in the State of Mexico (11.6%), Mexico City (7.3%), Puebla (6.6%), Jalisco (6.4%), and Veracruz (6%). But most of the workforce can be found in Mexico City (12.9%), State of Mexico (9.4%), Jalisco (6.9%), Nuevo León (5.7%), Guanajuato (5%), Veracruz (4.4%), and Puebla (4.4%). 36 out of 100 people work in any of these: State of Mexico, Mexico City, Jalisco or Nuevo León. This means that places with low in number of businesses but higher workforce are those with a lower proportion of micro-businesses.

So this research proposes to use this stereotype, and use it as an advantage for women in the financial industry. As their profile provide a better case study for financial portfolios.

Next section, we introduce the methods and strategies used to generate the woman profile.

3 Methods and Approaches

High dimensional data is problematic due to high computational costs and memory usage. It can consist of irrelevant, misleading or redundant features that are the ones increasing the search space size. They make it harder to process the data and therefore do not contribute to the learning process [13]. There are methods to reduce dimensionality: Feature Selection and Feature Extraction. In the former, the features that contain the information necessary to solve the problem are the only ones selected; in the latter, it consists of methods that seeks a transformation of the input space into a lower dimensional subspace that still contains most of the relevant information. Both or only one of these methods can

be used as a preprocessing step. The need for feature selection and extraction is due to the information of a dataset falling in either of three categories: relevant, irrelevant and redundant. By means of the method to remove the majority of irrelevant and redundant features, the objective is to create a subset with the least amount of features that still retains the most information and contributes to the learning accuracy [14].

Some advantages of feature selection [14] include:

- Reduce storage requirements and increase algorithm speed.
- Removes redundant, irrelevant or noisy data.
- Improves data quality.
- Increases accuracy of the resulting model.
- Reduction of feature set and therefore faster future recollections.
- Improved performance in predictive accuracy.
- Better visualization and understanding about the dataset.

There is a major difference between feature selection and feature extraction. In selection, when a dataset is complex with many relevant features, choosing only a small selection of them, means that information will be lost by the omission. Whereas in feature extraction, the size of the feature space can be decreased without discarding features, but the transformation done to linearly combine them makes them harder to interpret as individual features.

3.1 Principal Component Analysis

Among these techniques, the most common and popular one is Principal Component Analysis (PCA) developed by Karl Pearson in 1901. It consists of a linear transformation of data that minimizes the redundancy, which is measured by the covariance, and maximizes the information, which is measured by the variance. PCA uses orthogonal transformations to convert samples of correlated variables into samples of linearly uncorrelated features [13]. These new features are called principle components (PC) and describe the same or less amount of information. Principle components are new features with two properties:

- each principle component is a linear combination of the original variables.
- and the principle components are uncorrelated to each other and redundant information is discarded.

The principle components are ordered in terms of information gain. That is, the first principle component contains the most observed variability, and the last principle component contains the least. PCA reduces the number of original features by eliminating the last principle components.

The main applications of PCA are data compression, image analysis, visualization, pattern recognition, regression, and time series predictions.

PCA also has assumptions which limits its use [13]:

- relationship between variables is linear,
- variables are normalized,
- and it lacks a probabilistic model structure which is important in many statistical tests.

3.2 Unsupervised Clustering Techniques

K-Means. K-means is an unsupervised clustering algorithm whose objective is to assign data points into clusters, thus the overall distance between points and cluster centroids is minimized [8]. It tries to separate samples into groups of equal variance, minimizing a criterion known as inertia or within-cluster sum of squares [16]. Equation 1 shows the expression K-means minimizes:

$$\sum_{i=0}^n \min(\|x_i - \mu_j\|^2) \quad (1)$$

One of the advantages of this algorithm is that it scales well with data size and its flexibility allows it to be applied to many different fields.

The most common way to choose the number of clusters for K-means is using the elbow method. This is done by fitting K-means models for a range of consecutive numbers starting from 1, and plotting the total within sum of squares value for each number of clusters versus that cluster number. The elbow or turning point in the line plot, and therefore an adequate number of clusters, is where the reduction in sum of squares seems to drop off [8].

3.3 Machine Learning Algorithms

In this section, the techniques used will be discussed. Among the candidate methods for classification we used Decision Trees (DTs) which is a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Some advantages of decision trees are:

- Simple to understand and to interpret. Trees can be visualised.
- Requires little data preparation.
- The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree.
- Able to handle both numerical and categorical data.
- Able to handle multi-output problems.
- Uses a white box model.
- Possible to validate a model using statistical tests.
- Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.

The disadvantages of decision trees include:

- Decision-tree learners can create overfitting.
- Decision trees can be unstable.
- The problem of learning an optimal decision tree is known to be NP-complete under several aspects of optimality and even for simple concepts.
- There are concepts that are hard to learn because decision trees do not express them easily, such as XOR, parity or multiplexer problems.

- Decision tree learners create biased trees if some classes dominate. It is therefore recommended to balance the dataset prior to fitting with the decision tree.

3.4 Performance Measures

Accuracy is the most used measure to assess the performance of a credit evaluation model [7], and was seen throughout different studies [24,25] as shown in Eq. 2, 3 and 4.

$$\text{Type I accuracy} = \frac{\text{number of both observed bad and classified as bad}}{\text{number of observed bad}} \quad (2)$$

$$\text{Type II accuracy} = \frac{\text{number of both observed good and classified as good}}{\text{number of observed good}} \quad (3)$$

$$\text{Total accuracy} = \frac{\text{number of correct classification}}{\text{the number of evaluation sample}} \quad (4)$$

Other variations exist where the former measure was considered incomplete, such as in the study of Danenas and Garsvas [7]. Where the accuracy performance was represented using three different measures:

- Accuracy defined as the proportion of correct prediction shown in Eq. 5.

$$\text{accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total number of instances}} \quad (5)$$

- True Positive Rate (TPR) or Sensitivity is a ratio of True Positives over the total number of total positive instances shown in Eq. 6.

$$\text{TPR}_i = \frac{\text{True Positives}_i}{\text{True Positives}_i + \text{False Negatives}_i} \quad (6)$$

- F-Measure is defined as harmonic mean of precision and recall (True Positive) measures. It is preferred to accuracy for analysis of the classification performance case of unbalanced learning shown in Eq. 7.

$$F_i = \frac{2 \times \text{precision}_i \times \text{recall}_i}{\text{precision}_i + \text{recall}_i} \quad (7)$$

4 Results and Analysis

The population of interest in this research is focused on Mexican woman. Then, three sets of data were used in order to define the financial profile for this segment. The first dataset is from the Mexican National Institute of Statistics and Geography (INEGI). This data provide information about location and density of population in Mexico. The data was generated during the Population and Dwellings Census 2010 (ITER). A locality is any settlement where a community of people live. It can range from a small number of dwellings to a large urbanized

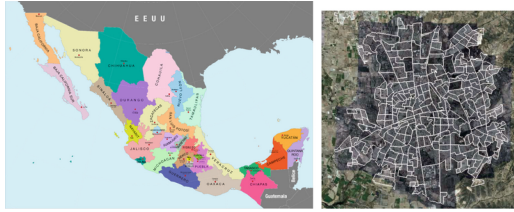


Fig. 1. Left: Mexican map with states. Right: Localities in a city

area. An example of how a locality looks like can be seen in Fig. 1. The dataset contains information on the localities for the thirty-two states found in Fig. 1.

The original dataset contained 196,025 localities and 200 attributes. The attributes contain demographic information of each locality. Usually in the form of counts or averages.

Among these, 23 attributes were chosen as relevant:

- 3 attributes related to coordinates
- 1 related to men-women ratio
- 1 related to child mortality
- 2 related to education
- 4 related to economic activity
- 3 related to civil status
- 2 related to household decisions
- 7 related to dwellings.

It can consist of irrelevant, misleading or redundant features that are the ones increasing the search space size. They make it harder to process the data and therefore do not contribute to the learning process.

4.1 Preprocessing

The programming language used was Python. Four steps were done to tidy and prepare the dataset for the next subsection:

- Coercing data types into floats
- Transforming attributes into proportions
- Scaling variables
- Removing nulls

By the end of the process, 93468 instances and 28 attributes remained. To reduce dimensionality, PCA was applied. 90% of the information was retained, and 11 principle components were created. The principle components percentages can be seen in Table 1. The weight of the variance for each principle component is of 0.2549, 0.2277, 0.1442, 0.0679, 0.0465, 0.0393, 0.0361, 0.0287, 0.0264, 0.0239, and 0.0200, respectively.

Table 1. Principle components percentage per attribute

PC	1	2	3	4	5	6	7	8	9	10	11
LONGITUDE	0.367	0.307	-0.371	-0.109	0.107	0.078	-0.026	0.008	-0.052	-0.103	-0.707
LATITUDE	0.389	0.492	-0.466	-0.060	-0.050	0.042	-0.021	-0.029	0.073	0.148	0.547
ALTITUDE	0.755	-0.633	0.048	0.047	-0.102	0.009	0.063	0.000	0.043	0.022	0.050
TAM.LOC	0.022	0.017	0.063	-0.071	-0.028	0.149	-0.200	0.946	-0.023	0.173	0.005
REL.H.M	0.000	0.000	-0.001	-0.001	-0.001	-0.001	-0.001	0.000	0.001	-0.001	0.000
PROM.HNV	-0.013	-0.039	-0.114	0.133	0.087	-0.015	0.029	0.017	-0.179	-0.067	-0.098
GRAPROES	0.126	0.192	0.361	-0.335	-0.120	0.018	0.132	-0.045	0.002	0.046	-0.042
GRAPROES_M	0.089	0.147	0.326	-0.311	-0.089	0.052	0.120	-0.036	-0.013	0.034	0.001
GRAPROES_F	0.154	0.220	0.358	-0.321	-0.142	-0.014	0.129	-0.050	0.011	0.051	-0.085
PEA_M	-0.011	0.085	-0.012	0.159	-0.229	-0.153	0.061	0.124	0.579	-0.376	-0.076
PEA_F	0.051	0.023	0.091	-0.093	-0.011	0.165	-0.026	0.029	0.051	-0.342	0.043
PE_INAC_M	0.017	0.021	0.035	-0.034	0.181	0.036	0.098	-0.060	-0.176	0.399	-0.012
PE_INAC_F	-0.020	0.009	0.009	0.208	0.187	-0.170	0.228	0.010	0.100	0.508	-0.147
P12YM_SOLT	-0.001	-0.020	0.031	-0.110	0.373	-0.111	0.142	0.030	0.681	0.189	-0.027
P12YM_CASA	0.034	0.128	0.069	0.289	-0.391	-0.174	0.090	0.048	-0.143	0.086	-0.194
P12YM_SEPA	0.006	0.042	0.027	0.097	0.197	0.237	0.175	0.041	-0.002	-0.151	0.039
HOGJEF_M	0.019	0.118	0.038	0.281	-0.311	-0.062	0.104	0.021	0.067	0.075	-0.059
HOGJEF_F	0.036	0.031	0.088	0.113	0.414	0.442	0.288	0.051	-0.083	-0.236	-0.005
PROM_OCUP	-0.027	-0.089	-0.067	-0.200	0.055	-0.123	-0.162	-0.032	-0.007	0.040	0.042
PRO_OCUP_C	-0.015	-0.016	-0.068	-0.032	-0.042	0.001	-0.003	-0.007	0.028	-0.024	-0.019
VPH_PISODT	0.134	0.198	0.234	0.387	-0.010	0.116	0.234	0.026	-0.045	-0.020	0.253
VPH_1CUART	-0.045	0.011	-0.028	0.000	-0.093	0.118	0.005	0.000	0.024	-0.103	0.090
VPH_2CUART	-0.109	-0.065	-0.090	0.122	-0.316	0.596	0.168	-0.095	0.190	0.275	-0.163
VPH_3YMASC	0.170	0.175	0.192	0.268	0.261	-0.388	0.143	0.125	-0.146	-0.150	-0.038
VPH_C.SERV	0.167	0.174	0.355	0.308	0.152	0.199	-0.746	-0.208	0.147	0.112	-0.072

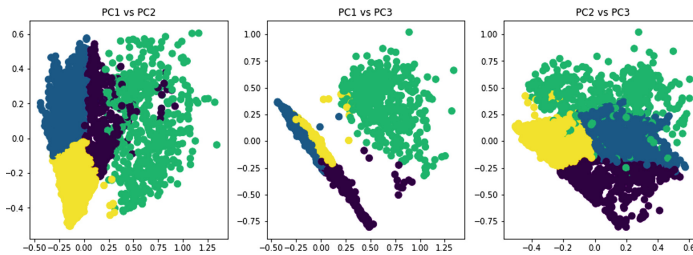


Fig. 2. PCA plot

Clustering. To identify the best number of clusters found in the data, three different scores were used: the Elbow method, the Davies-Bouldin score and the Calinski-Harabasz score. The results obtained in Fig. 3 show that four clusters was the best option overall. In the elbow method, it is were the distances start decaying; in Davies-Bouldin score is the lowest score; and in Calinski-Harabasz score is the highest value. Applying the Silhouette score gives us a value of 0.377.

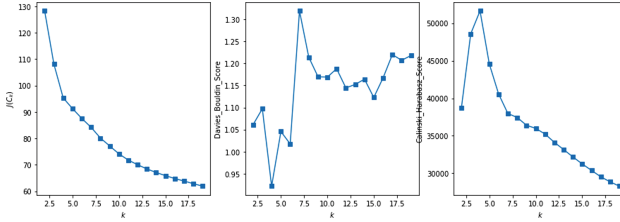


Fig. 3. Cluster scores

The reduced data then is used for clustering the population. Results show that the country is divided into four clusters as Fig. 4 shows: a North cluster (blue), a South cluster (red), a Central cluster (cyan) and Small populations clusters (yellow). By seeing them individually, such as in Fig. 5, we can appreciate the size of each cluster without the overlapping. As such, the Small Populations cluster is a very big cluster, which defining characteristic is the small size of the population in the localities. And contrary to the other clusters found, this can be found all over the country.

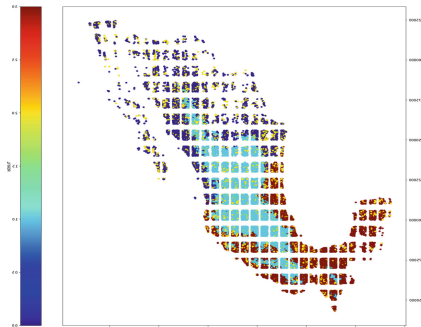


Fig. 4. Clusters for Mexican population (Color figure online)

But as we are dealing with proportions, it gives very inflated values as its population size is very small. Which also make the populations with very large populations have smaller proportions overall, and thus is excluded for further analysis to concentrate on the bigger localities.

An interesting trend found was within the proportion of household heads found in the country found in Fig. 5, where the proportion of household heads and the Economically Active Population per Locality are plotted. It is shown that the household authority of men grow proportionally with their economic activity. Whereas the household authority of women decreases with theirs. Women as mentioned in literature, are not as economically active as men, but when the economic activity of women increases the trend seems to grow positively.

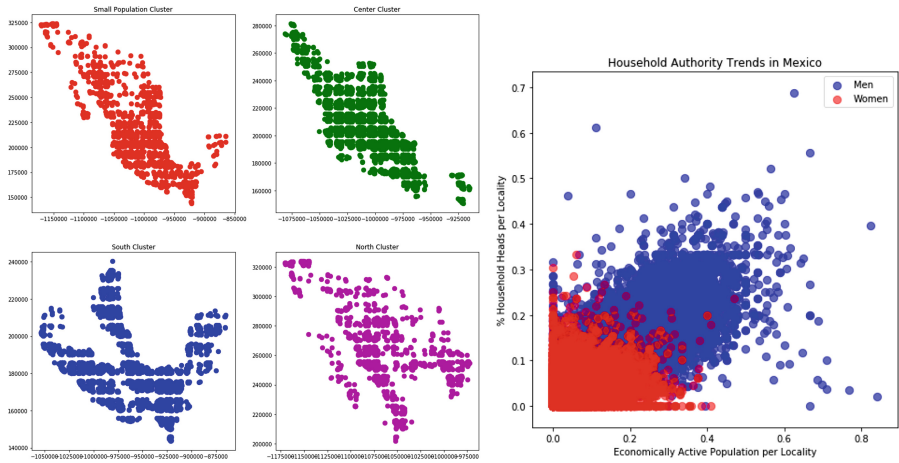


Fig. 5. Left: individual clusters. Right: household authority in Mexico (Color figure online)

Another trend can be seen in married people per locality and economic activity. The economic activity of men increases as the percentage of married men increases per locality. The case of women is a bit trickier though as shown in Fig. 6. As the economic activity of women increases, the proportion of married women per locality stays averaged approximately in 40%.

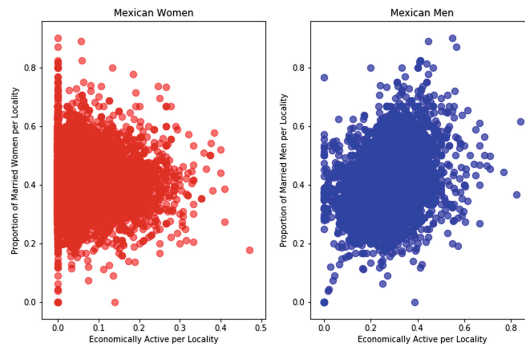


Fig. 6. Civil status in Mexico

On the other hand, the average academic grade per locality is pretty even between genders, just as mentioned in literature about LAC countries. A linear regression shows that both levels are similar just as shown in Fig. 7.

Decision Trees. The decision tree was created using an implementation found in scikit-learn. The inputs were the attributed without PCA reduction and the

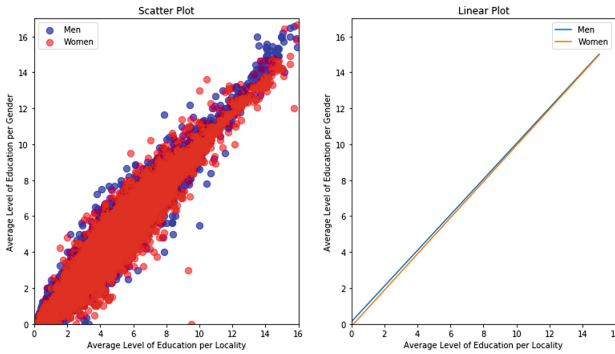


Fig. 7. Education level per locality

clusters obtained during K-Means. Once again the cluster with small populations was omitted.

A training and testing set of 70% and 30%, respectively was used. The parameters used were a criterion of entropy, a max depth of 5 and minimum samples per leaf of 30. The score obtained was of 0.955.

In most cases, the coordinates and altitude of the locality where used to separate the clusters. But in cases where overlapping existed between clusters, variables such as the level of education and percentage of houses with more than three rooms were the deciding factors. The North cluster has a higher level of education per locality and houses with more than three rooms than the other clusters, the Center cluster has a higher proportion as well than the South cluster.

We can observe localities are defined by their location, altitude, average level of education and the percentage of houses with more than three rooms. The North cluster was found to have a higher level of education and number of rooms than the Center and South cluster where overlapping existed. The Center Cluster had as well a higher level of education and rooms than the South cluster.

Many trends found in literature could also be found in the data such as an even education, a smaller proportion of female household heads and some trends with economic activity and civil status between genders.

The Small Populations cluster is another interesting result. It was the only cluster that was not defined strongly by its geography and could be found around all the country. Still, the small sample size brought along many overestimated proportions and could not be used to compare against the clusters with larger localities.

The results provide more knowledge on the partition of the Mexican population using state-of-the-art Machine Learning techniques. Prior studies made using census data on the Mexican population do not make use of them, whereas other countries [1, 4, 19, 20] have started implementing them on their own data with good results.

5 Conclusion

This research provide a methodology for analyzing several datasets using machine learning techniques in order to define a financial profile of Mexican Woman. The methodology includes first the preprocessing of the data, to then reduce the dimensionality using a PCA. Followed by a clustering analysis, and finally the use of decision trees to define the profile. The main contribution of this research is the definition of the mexican financial profile for women segment. Highlighting that women segment has been partially attended by the financial entities, missing the opportunity to offer financial products according to the risk involved. The methods applied in this research have been proof to be more flexible and accurate than traditional statistical approaches. Future work includes the risk calculator for specific financial products, as well as the extension for population of other countries of interest.

References

1. Abarca-Alvarez, F., Campos, S., Bellido, R.: Señales de gentrificación a través de la inteligencia artificial: identificación mediante el censo de vivienda. *Bitácora Urbano Territorial* **28**, 103–114 (2018)
2. World Bank. The effect of women's economic power in Latin America and the Caribbean. Latin America and the Caribbean poverty and labor brief, p. 48 (2012)
3. Bertocchi, G., Torricelli, C., Brunetti, M.: Portfolio choices, gender and marital status. *Rivista di Politica Economica* **98**, 119–154 (2008)
4. Sheng, B., Gengxin, S.: Data mining in census data with cart. In: 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), vol. 3, pp. V3-260–V3-264, August 2010
5. Brindley, C.: Barriers to women achieving their entrepreneurial potential: women and risk. *Int. J. Entrepreneurial Behav. Res.* **11**(2), 144–161 (2005)
6. Charness, G., Gneezy, U.: Strong evidence for gender differences in risk taking. *J. Econ. Behav. Organ.* **83**(1), 50–58 (2012). Gender Differences in Risk Aversion and Competition
7. Danenas, P., Garsva, G.: Selection of support vector machines based classifiers for credit risk domain. *Expert Syst. Appl.* **42**(6), 3194–3204 (2015)
8. Flynt, A., Dean, N.: DA survey of popular R packages for cluster analysis. *J. Educ. Behav. Stat.* **41**, 205–225 (2016)
9. Hachicha, N., Jarboui, B., Siarry, P.: A fuzzy logic control using a differential evolution algorithm aimed at modelling the financial market dynamics. *Inf. Sci.* **181**(1), 79–91 (2011)
10. Hsieh, N.-C.: An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert Syst. Appl.* **27**(4), 623–633 (2004)
11. INEGI. Mujeres y hombres en México 2016. Instituto Nacional de Estadística y Geografía (México), p. 250 (2016)
12. Jenny, S.-S.: Self-directed pensions: gender, risk, and portfolio choices. *Scand. J. Econ.* **114**(3), 705 (2012)
13. Khalid, S., Khalil, T., Nasreen, S.: A survey of feature selection and feature extraction techniques in machine learning. In 2014 Science and Information Conference, pp. 372–378, August 2014

14. Ladla, L., Deepa, T.: Feature selection methods and algorithms. *Int. J. Comput. Sci. Eng. (IJCSE)* **3**, 1787–1797 (2011)
15. Masthead. Factors that every financial advisor must consider when doing a client's risk profile, September 2016. <https://www.masthead.co.za/newsletter/factors-that-every-financial-advisor-must-consider-when-doing-a-clients-risk-profile/>
16. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
17. Petreska, B.R., Kolemisevska-Gugulovska, T.D.: A fuzzy rate-of-return based model for portfolio selection and risk estimation, pp. 1871–1877, October 2010
18. Rybczynski, K.: Gender differences in portfolio risk across birth cohort and marital status. *Can. J. Econ./Revue canadienne d'économique* **48**(1), 28–63 (2015)
19. Shanmuganathan, S., Li, Y.: An AI based approach to multiple census data analysis for feature selection. *J. Intell. Fuzzy Syst.* **31**, 859–872 (2016)
20. Sharath, R., Krishna Chaitanya, S., Nirupam, K.N., Sowmya, B.J., Srinivasa, K.G.: Data analytics to predict the income and economic hierarchy on census data. In 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), pp. 249–254, October 2016
21. Simms, K.: Investor profiles: meaningful differences in women's use of investment advice? *Financ. Serv. Rev.* **23**(3), 273–286 (2014)
22. Thomas, L.C.: A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *Int. J. Forecast.* **16**(2), 149–172 (2000)
23. UN. Economic Empowerment of Women. Technical report, United Nations (2013)
24. Yu, L., Wang, S., Lai, K.K.: Credit risk assessment with a multistage neural network ensemble learning approach. *Expert Syst. Appl.* **34**(2), 1434–1444 (2008)
25. Yu, L., Yue, W., Wang, S., Lai, K.K.: Support vector machine based multiagent ensemble learning for credit risk evaluation. *Expert Syst. Appl.* **37**(2), 1351–1360 (2010)
26. Zavala-Diaz, J.C., Díaz-Parra, O., Ruiz-Vanoye, J.: Analysis of risk in linear multi-objective model and its evaluation for selection of a portfolio of investment in the Mexican stock exchange. *Afr. J. Bus. Manage.* **5**, 7876 (2011)