ISSN 2277-3061

# Discovering Mexican Birth Rate Patterns Using Machine Learning Techniques

Maria Somodevilla, David Limón, Ivo Pineda, Concepción Pérez de Celis, Darnes Vilariño
Autonomous University Benemerita of Puebla, Mexico
mariajsomodevilla@gmail.com
Autonomous University Benemerita of Puebla, Mexico
david_d_93@hotmail.com
Autonomous University Benemerita of Puebla, Mexico
ivopinedatorres@gmail.com
Autonomous University Benemerita of Puebla, Mexico
mcpcelish@gmail.com
Autonomous University Benemerita of Puebla, Mexico
dvilarinoayala@gmail.com

## ABSTRACT

In this research, we attempt to discover patterns that describe and predict the birth rate in Mexico by using data mining techniques based on relevant demographic and economic information about Mexico. More than twelve million births data obtained from the General Directorate of Health Information in the period 2008-2013 were analyzed. The acquired knowledge allows us to say that in Mexico the birth rate is affected by the social welfare, education and marginality at county level. Due to the diversity of the population and the large number of socioeconomic factors involved in Mexican society, it is difficult to find general impact factors for this issue. The results of this research are not intended to be definitive but its aim is to provide indicators that may influence decisions about birth control in Mexico.

## Indexing terms/Keywords

Data Mining; Birth Rate; Mexico; Clustering; Classification.

## Academic Discipline And Sub-Disciplines

Computer Science, Data Mining, Machine Learning

## SUBJECT CLASSIFICATION

Data Mining Classification

## TYPE (METHOD/APPROACH)

Descriptive analysis, Predictive analysis, SimpleKMeans, J48 algorithm

Council for Innovative Research

Peer Review Research Publishing System

Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY

Vol. 15, No. 1

www.ijctonline.com, editorijctonline@gmail.com

## INTRODUCTION

In Mexico, the population has shown an increasing trend in recent years. According to the population census conducted by INEGI (National Institute of Statistics and Geography) this increase has been constant, with an average of 7 771.723 inhabitants per year [1], as shown in Figure 1.

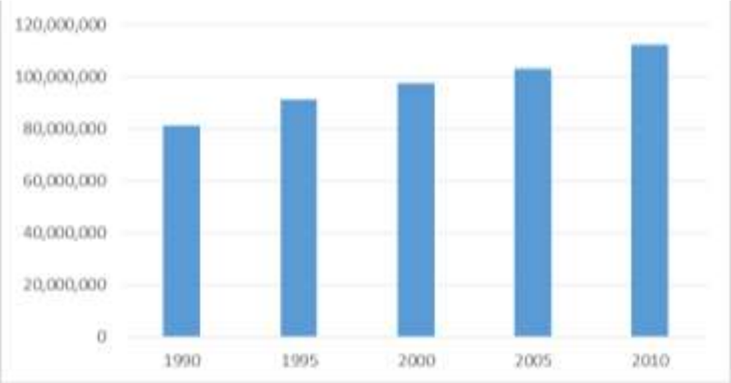

**Fig. 1: Population growth trends (2008 – 2013)**

There are various socio-economic and demographic factors that influence birth. Because of the diversity and density of each state population, these factors vary and are difficult to determine. Analyzing more than twelve million births data [2] obtained from the DGIS (Directorate General of Health Information) in the period 2008-2013 some growth trends were found, as well as periods in which the trend has been downward.

## 1.1 Reduction in births

The year 2013 showed the first sign of decreasing in births number over the previous year (2012, see Figure 2).

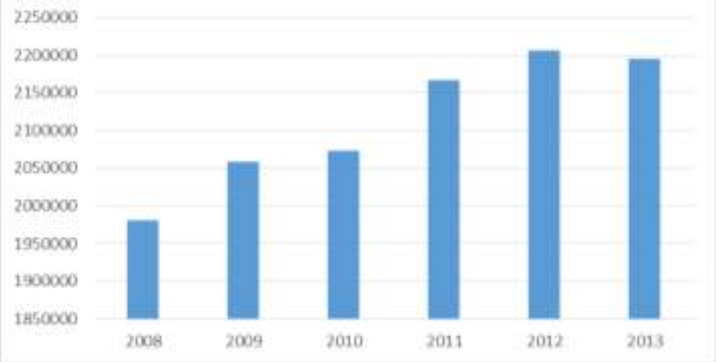

**Fig. 2: Number of births Mexico 2008 – 2013**

Similarly, the graph in Figure 3 shows a decrease in the number of births in the last three years above (2011, 2012 y 2013).

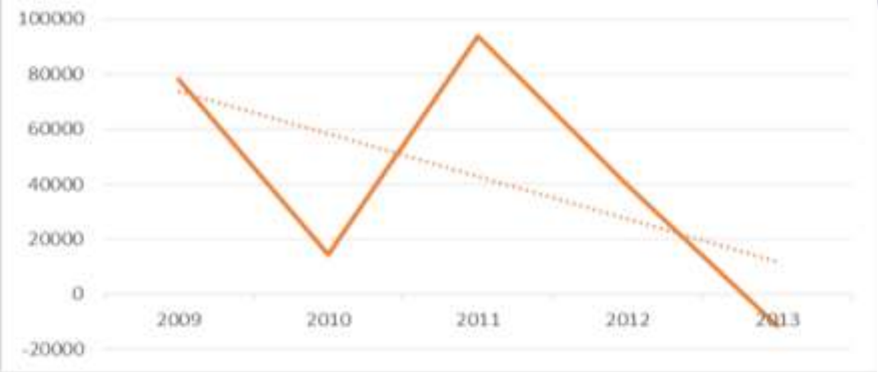

**Fig. 3: Variation of births compared to the previous year**

The births annual growth rate has fallen to have negative numbers, which occurs similarly in other countries. According to a similar investigation by the Planning Board the number of live births in Puerto Rico has declined significantly in recent decades [3]. Note that in the 2010-2011 biennials a large increase in births was documented. Although in general the number of births has been increasing every year, it is important to note this increase does not determine the trend, since the birth rate continues to decline in future years, the births number will go down too.

## 1.2 Births by Mexican states

Looking at the data graphically represented in the political-administrative map of the Mexican Republic in Fig. 4, we can observe the annual increase of births by states.

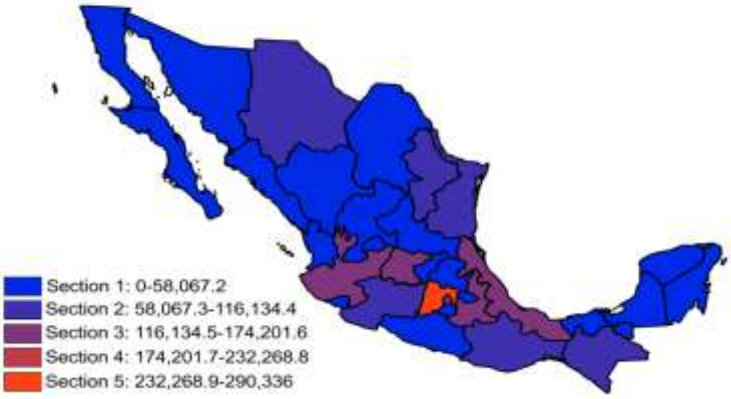
Section 1: 0-58,067.2
Section 2: 58,067.3-116,134.4
Section 3: 116,134.5-174,201.6
Section 4: 174,201.7-232,268.8
Section 5: 232,268.9-290,336

**Fig. 4: Births in 2010 by state, colors: cooler-fewer births, warmer-a great deal**

Notice the central-southern Mexico has a greater number of births. It is important to consider the population density of each state, which is why the Federal District and state of Mexico have a noticeably warmer color.

## DATA DESCRIPTION

Data was obtained from DGIS in .xls format; approximately 17% of them were missing. Attributes such as p_con_derecho_ss(with social welfare), p_sin_derecho_ss(without social welfare), p_con derecho_ss_h (men with social welfare), p_sin_derecho_ss_h (men without social welfare) and mpo_nacim (county of birth) showed no variation in their values and therefore not considered significant and were eliminated. After debugging 139.793 rows and 25 columns remain. The complexity of this work lies in the dataset nature. Since data come from all Mexican states, it is difficult to make an analysis without involving other factors which are scarce or nonexistent. A description of the attributes is presented in Table 1.

**Table 1. Dataset attributes.**

| Attribute | Data type | Mean | Standard deviation |
|---|---|---|---|
| edadmadre (mother age) | numeric | 26.06 | 6.309 |
| estado_civil (mother marital status) | nominal | --- | --- |
| numero_embarazos (# of pregnancies ) | numeric | 2.258 | 1.433 |
| nacidos_muertos (# of stillbirths) | numeric | 0.151 | 0.456 |
| nacidos_vivos (# of live births) | numeric | 2.109 | 1.317 |
| atencion_prena (prenatal care received) | nominal | -- | -- |
| consultas (# of visits to the doctor) | numeric | 6.998 | 3.259 |
| Derechohabiencia (social welfare) | nominal | -- | -- |
| escolaridad (education) | nominal | -- | -- |
| ocupacion_habitual (mother occupation) | nominal | -- | -- |
| desc_ocuphab(mother occupation descript) | nominal | -- | -- |
| sexo_rn (newborn sex) | nominal | -- | -- |
| gestach (gestational age in weeks) | numeric | 38.81 | 1.817 |
| tallah (newborn size) | numeric | 49.92 | 3.159 |
| pesoh (newborn weight in grams) | numeric | 3139.49 | 525.954 |
| procedimiento (delivery method) | nominal | -- | -- |

| Attribute | Data type | Mean | Standard deviation |
|---|---|---|---|
| atendio_parto (person delivering the baby) | nominal | -- | -- |
| lugar_nacim (childbirth place) | nominal | -- | -- |
| índice_marginacion_municipio (county marginalization rate ) | numeric | 14.044 | 8.956 |
| tasa_bruta_natalidad (county  birth rate) | nominal | -- | -- |
| defunciones_generales(# general deaths) | numeric | 3086.31 | 4292.877 |
| p_sin_derecho_ss_m (county population without welfare) | numeric | 76171.7 | 89297.957 |
| p_con_derecho_ss_m (county population with welfare) | numeric | 159919.6 | 163054.064 |
| p_con_derecho_privada_m (county population with private welfare) | numeric | 6136.39 | 8915.058 |

## DATA PREPROCESSING

For this analysis, births data from January 2010 was used, as well as demographic and socio-economic data listed in section 2. Once removed missing values, the data set contains 139,793 minable instances. Figure 5 is a summary of the birth quantity variation in a six year period, including the year of analysis.

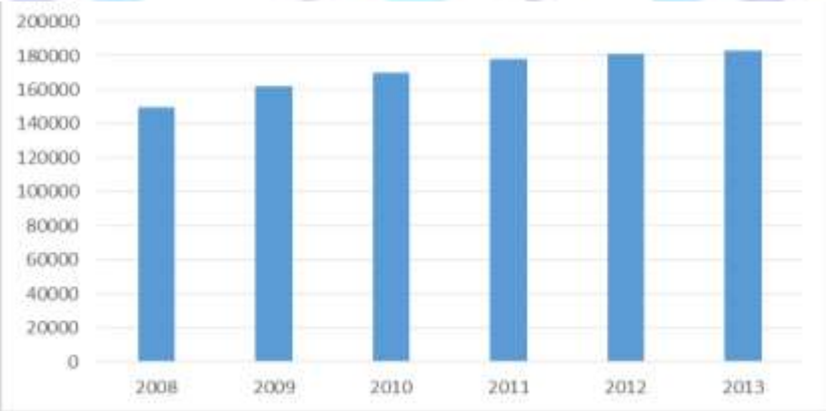

**Fig. 5: Birth quantity in January, 2008 – 2013**

Since January contains less data variability and a growth trend is comparable to the overall sample, then January was considered the most feasible to conduct an analysis to return some degree of generalizable knowledge.

## DATA MINING APPLICATION

The data mining process proposed consist of performing a mother´s data grouping according to common characteristics. This minable view was also applied predictive techniques to classify considering the birth rate and find their relationship with socio-economic and demographic factors.

### 4.1 Descriptive analysis

The objective of this analysis is to generate groups with similar characteristics that will be later used as classes. In particular, it is in our interest to analyze the attributes that provide meaningful information about mothers, and thus generate statistical estimators to segment the population. Due to the volume and variety of data, the clustering algorithm was carried out with different group sizes. The decision to define 5 groups for the grouping is based on the fact of reaching a balance between error and the data separation. An excessive segmentation results in too many classes to perform effective analysis, and a low segmentation, which would not provide enough diversity. The SimpleKMeans algorithm was selected because of its simplicity and effectiveness.

Due to the characteristics of the data set used for this analysis, it is necessary to find the best way to identify groups. The groups that are identified may be exclusive so that any instance belongs in only one group or they may be overlapping. They may be probabilistic, whereby an instance belongs to each group with a certain probability or they also may be hierarchical. The final clusters are quite sensitive to the initial cluster centers. It is almost always infeasible to find globally optimal clusters. To increase the chance of finding a global minimum people often run the algorithm several times with different initial choices and choose the best final result, the one with the smallest total squared distance [4].  To achieve this goal 10 experiments with the data set using different values for the seed were performed.  Tests with different values for k yielded different errors as shown in Table 2. Note that the best grouping was obtained with five clusters with a mean square error of 333,813.67.

Table 2. Grouping tests performed with different values for K

| Number of clusters (k) | Error |
|---|---|
| 2 | 483002.936 |
| 4 | 444813.643 |
| **5** | **333813.67** |
| 6 | 424179.276 |
| 8 | 602722.11 |
| 10 | 389036.982 |

The features distinguishing the cluster#0 are: professional women (Fig. 6) of about 28 years old with social welfare, no paid employment in counties with low exclusion rate (9,922), intermediate birth rate (22-33 births per 1000 inhabitants), large number of deaths (4734) and about 1 in 3 women will not have any type of welfare.

Given the extensive range of numbers represented in the class color gradient, is it difficult to observe notable differences in the cluster assignment seen at Fig. 7. However, it is notable; the color ranges in the middle of the color class gradient. Also in Fig. 8 the left side has a notable difference in color and many more yellow spots, which tells us that the assignment of that specific data section is very diverse and has a higher welfare.

Cluster#1 comprises women of about 26 years old with popular welfare and complete secondary which gave birth through normal delivery in municipalities with a higher rate of marginalization and low birth rate. As can be noted in Fig.6, the predominant color is dark blue as most women having children have a middle school, degree of school.

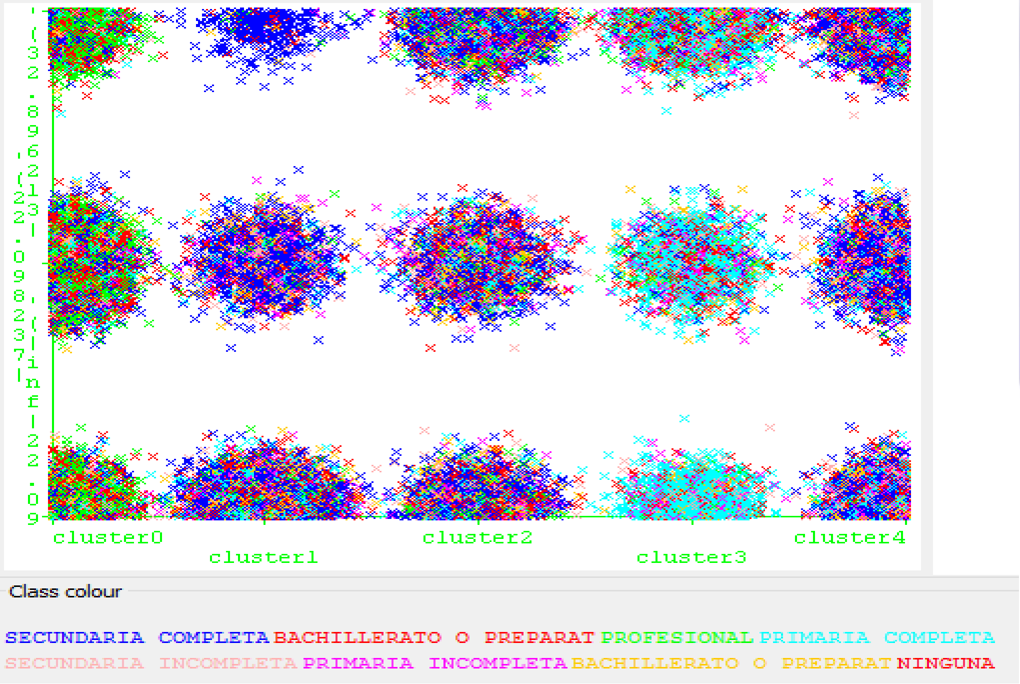

**Fig. 6: Cluster assignments. X axis: cluster, Y axis: birth rate, colors: education**

Cluster#2 is characterized by women about 25 years old in free union , without any welfare and gainful occupation, marginalization index average (13.83), and high birth rate (33 births over 1,000 inhabitants). As it can seen in Fig. 7, the predominant class color is a blue darker tone which means a lower to medium marginalization degree.

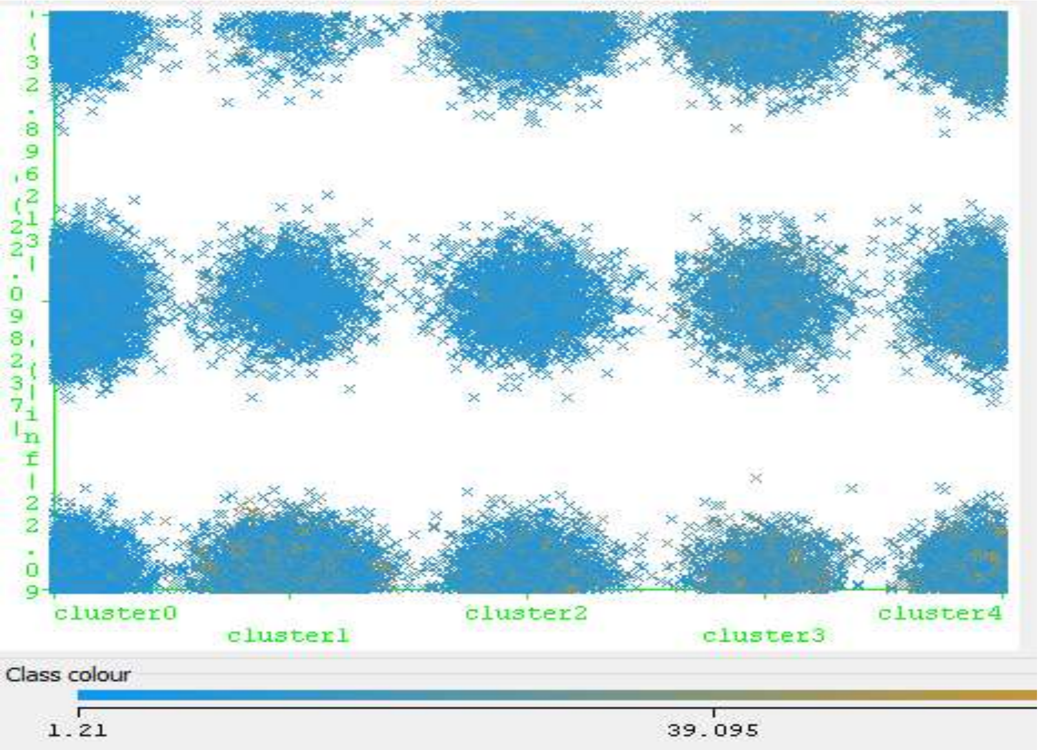

**Fig. 7: Cluster assignments. X axis: cluster, Y axis: birth rate, colors: marginalization**

Cluster#3 is composed of about 27 year old women with popular welfare, elementary education (Fig. 6), intermediate marginalization (13.83) and a high birth rate (33 births or more per 1000 inhabitants). Observing Fig. 8, it can be seen the cluster assignments and the predominant darker blue tone class color, (few women with social welfare) in the outer right cluster assignments, whereas the left cluster assignments have a significant amount of blue-green combination color which means many women with social welfare.

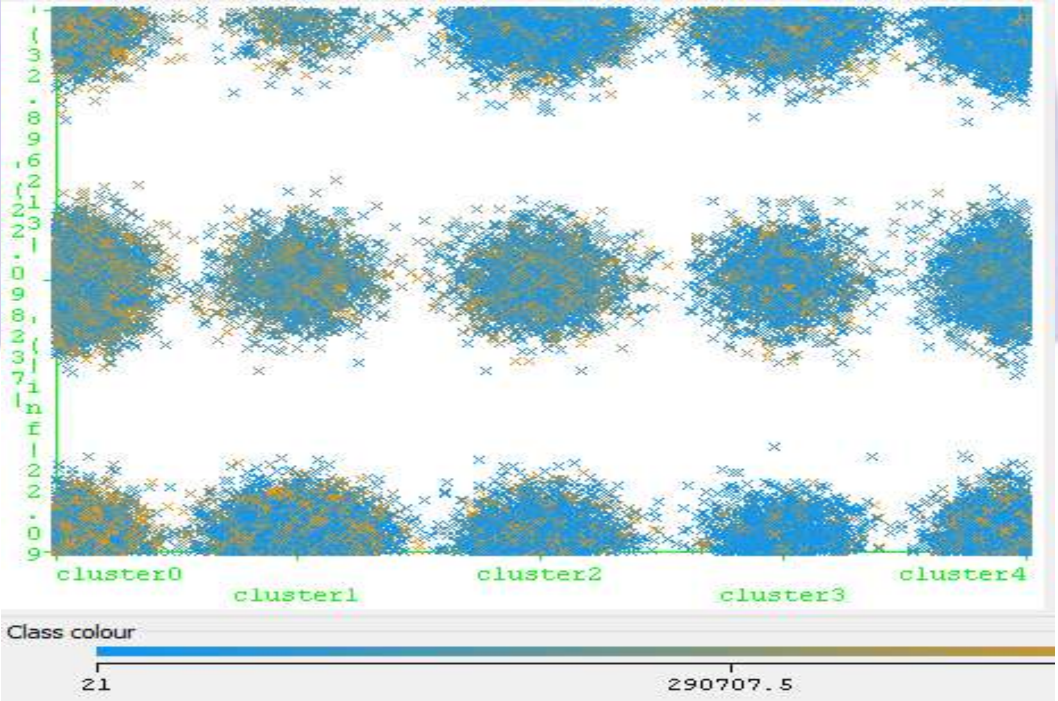

**Fig. 8: Cluster assignments. X axis: cluster, Y axis: birth rate, colors: welfare**

Finally, the cluster#4 is characterized by 24 year old women in free union and popular welfare, high rate of marginalization (16.23) and a high birth rate (33 births per 1,000 or more inhabitants). Detailed statistical information for clusters is shown in Table 3.

Table 3. K-means clusters description.

| Attribute | full data | cluster#0 | cluster#1 | cluster#2 | cluster#3 | cluster#4 |
|---|---|---|---|---|---|---|
| Instances | 92263 | 21318 | 16409 | 18056 | 13986 | 22494 |
| mother age | 26.0596 | 28.2692 | 25.8678 | 25.1453 | 27.0894 | 24.1989 |
| marital status | married | married | married | unión libre | married | free union |
| #of pregnancies | 2.2545 | 2.099 | 2.4252 | 2.0289 | 2.703 | 2.1797 |
| # of stillbirths | 0.1518 | 0.1874 | 0.1522 | 0.1398 | 0.1683 | 0.1171 |
| # of live births | 2.1042 | 1.9181 | 2.2731 | 1.8917 | 2.5333 | 2.0609 |
| prenatal care | yes | yes | yes | yes | yes | yes |
| #visits to doctor | 6.9985 | 8.4693 | 6.529 | 6.9659 | 6.5004 | 6.2828 |
| social welfare | popular | imss[1] | popular | nothing | popular | popular |
| education | middle | profesional | middle | middle | elementary | middle |
| mother_occup. | no_paid | no_paid | no_paid | no paid | no_paid | no_paid |
| occup_ descrip | housewife | housewife | housewife | housewife | Hogar | housewife |
| newborn sex | man | woman | man | man | man | woman |
| gestation(weeks) | 38.8121 | 38.5352 | 38.9096 | 38.6932 | 38.9406 | 39.0191 |
| newborn size | 49.917 | 49.7398 | 50.2362 | 49.7666 | 50.0864 | 49.8677 |
| newborn weigth | 3138.95 | 3131.95 | 3196.40 | 3123.80 | 3168.86 | 3097.25 |
| delivery method | normal | cesarean | normal | cesarean | normal | normal |
| delivering baby | M.D. | M.D. | M.D. | M.D. | M.D. | M.D. |
| childbirth place | state-ruled hospital | imss | state-ruled hospital | state-ruled hospital | state-ruled hospital | state-ruled hospital |
| marginalization | 14.0377 | 9.922 | 15.0459 | 13.8393 | 15.8457 | 16.238 |
| birth rate | '(22.098-32.896)' | '(22.098-32.896)' | '(-inf-22.098)' | '(32.896-inf)' | '(32.896-inf)' | '(32.896-inf)' |
| #general deaths | 3101.488 | 4733.910 | 2768.688 | 3072.8077 | 2835.6148 | 1985.517 |
| Pop. w/out welfare | 76220.280 | 107180.44 | 89772.907 | 74270.23 | 59152.464 | 49169.806 |
| pop. with welfare | 160185.86 | 235075.16 | 187730.46 | 143717.98 | 124354.20 | 104616.23 |
| population with private welfare | 6158.3971 | 9549.393 | 7232.8117 | 5848.9824 | 4169.1389 | 3646.1376 |

In general, the groups found tell us that there is some relationship between the marginality rate, education, birth rate, deaths and social welfare, which are the factors that show a greater variation between groups. It was also confirmed that those groups with more schooling have a lower birth rate.

## 4.2 Predictive analysis

Decision trees are the most powerful approaches in knowledge discovery and data mining. It includes the technology of research large and complex bulk of data in order to discover useful patterns [5].

For predictive analysis, J48 algorithm is used, which generates a classification tree. A classification tree is used to learn a classification function which concludes the value of a dependent attribute given the values of the independent attributes [5]. Decision makers prefer a decision tree because it is not complex as well as easy to understand. Tree complexity has

[1] Mexican Social Security Institute

its effect on its accuracy. J48 was applied to the same data set used for clustering in section 4.1, with the difference of having discretized birth rate in three baskets with equal frequency, representing the three birth rates low, medium and high.

The tree obtained is shown in Figure 9. The size of the tree is 47 with a total of 24 leaves and 4000 instances. Correctly classified instances amounted from 51% to 71.56%. To achieve this percentage cross-validation was performed with the aim of increasing the effectiveness of the development model based on the training data to classify more accurately later. This method has a high computational cost which is not recommended for all types of analyzes, especially those who do not suffer from the curse of dimensionality. The attributes considered for classification were deaths, social welfare and marginalization rate since they were the ones who had greater weight to predict the birth rate according to an algorithm of principal components.  The leaves have no more children represent the conclusion of the class variable (birthrate).

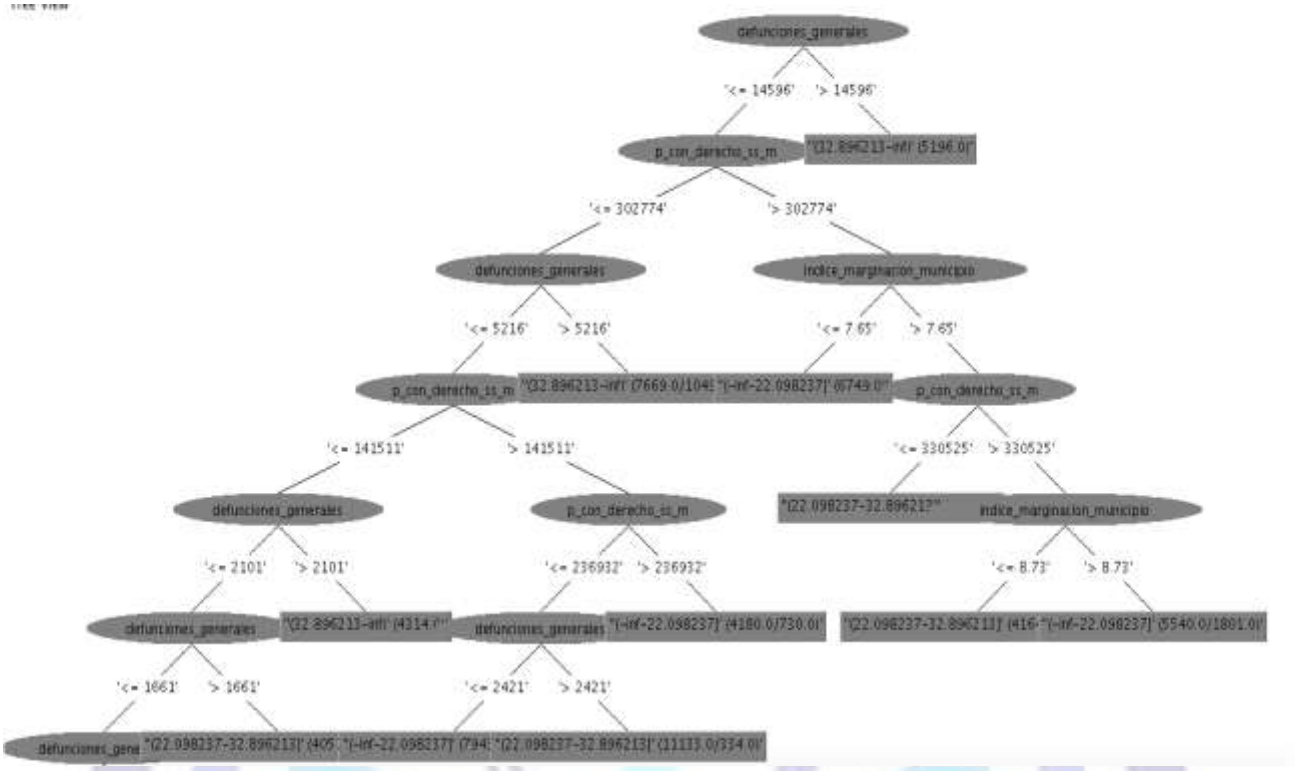

**Fig. 9: Simplified J48 classification tree, trimmed last 12 levels**

Since decision trees are one of the most powerful methodologies for autonomous learning, its use in this case is quite adequate, especially the ability to analyze on large and complex data sets to uncover patterns useful and interesting. The confusion matrix (see Table 4) indicates the largest number of instances well classified fall under the range of high birthrate (32.89 or greater), however,  instances not lying on the main diagonal represent a considerable part of the dataset. A result close to 75% correctly classified instances is acceptable to say that these results are relevant taking into consideration the issues discussed in Section 2.

**Table 4.  Confusion matrix generated by J48 classifier.**

| a | b | c | classified as |
|---|---|---|---|
| 32779 | 9349 | 4424 | a = '(-inf-22.098237]' |
| 8406 | 29854 | 8587 | b = '(22.098237-32.896213]' |
| 5947 | 3044 | 37403 | c = '(32.896213-inf)' |

## CONCLUSIONS

In this research the characteristics of births in the Mexican Republic were analyzed, subsequently machine learning techniques were applied to find patterns and outstanding features in the data set. First of all, a descriptive analysis in which groups that describe segments of the general population generated was performed. Afterwards, a predictive analysis in which instances are classified using a classification tree algorithm J48 was carried out. The decision tree shows municipalities with high rates of overall deaths exhibiting a high birth rate, but when the death number is lower (populations with smaller populations) rules characterizing the birth rate become more complex, analyzing then social welfare and municipal marginalization index. The acquired knowledge allows us to say that in Mexico the birth rate is affected by the social welfare, education and marginalization at the county level. Due to the diversity of the population and the large number of socioeconomic factors involved in Mexican society, it is difficult to find general impact factors for this issue. The results of this research are not intended to be definitive but its aim is to provide indicators that may influence decisions about birth control in Mexico.

## ONGOING WORK

It is working on comparing the results of this study with those obtained in other countries using the NoSQL data analysis. This could improve the process of obtaining results in queries to leverage the efficiency and management of large volumes of data characteristic of this management system database.

## ACKNOWLEDGEMENTS

We thank the Vicerector for Researcher and Postgraduate Studies at the Autonomous University of Puebla for supporting this work.

## REFERENCES

[1] Tendencias de crecimiento poblacional Mexico 2008 – 2013, http:// www.inegi.org.mx/lib/olap/consulta/general_ver4/MDXQueryDatos.asp?proy=sh_pty5ds [Online 07 2015].

[2] Estadísticas de nacimientos Mexico 2008 – 2013, http://www.dgis.salud.gob.mx/contenidos/basesdedatos/da_index.html [Online 07 2015].

[3] L. G. Pelatti: Suplemento Especial: Natalidad. San Juan, Puerto Rico (2013).

[4] I. H. Witten and E. Frank: Data Mining Practical Machine Learning Tools and Techniques, Third Edition, Morgan Kaufmann Publisher: Elsevier Inc., (2011).

[5] N. Bhargava, G. Sharma, R. Bhargava and M. Mathuria: Decision Tree Analysis on J48 Algorithm for Data Mining. International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, issue 6, 1114-1119 (2013).

[6] The World Bank Group, Birth rate crude (2015), http:// http://data.worldbank.org/indicator/SP.DYN.CBRT.IN/countries

[7] W. Fan and A. Bifet: Mining Big Data: Current Status, and Forecast to the Future. SIGKDD Explorations, vol. 4, issue 2(2014).