TEST Engineering & Management

May – June 2020
ISSN: 0193-4120 Page No. 17937 - 17946

Bayesian Model to Determine Genealogical Links of Family Descendants

Ciro Rodriguez[1], Pedro Lezama[2], Freddy Kaseng[3], Danilo Chávez[4]

[1,2,3]National University Federico Villarreal, Lima, Peru
[4,5]National Register of Identification and Civil Status RENIEC
[1]crodriguez@unfv.edu.pe, [2]plezama@unfv.edu.pe, [3]fkaseng@unfv.edu.pe fjimenez@reniec.gob.pe,
[4]dachavez@reniec.gob.pe

Article Info
Volume 83
Page Number: 17937 - 17946
Publication Issue:
May - June 2020

Article History
Article Received*: 1May 2020*
Revised*: 11 May 2020*
Accepted*: 20 May 2020*
Publication*: 24May 2020*

Abstract

Knowing our family ancestry and descendant is essential since we are the result of generations that have transmitted lives and experiences over time; however, today, it is also of great help in genetics, health, and safety, among other fields. The research shows the development of a model for the genealogical linking of citizens by a descendant, using Bayesian methods with machine learning algorithms, processing data from structured sources of civil registries, and identification of Peru as RENIEC. The model verifies the relationship and identity of citizens' links by determining their descent by integrating and analyzing incomplete, incorrect, inaccurate, and irrelevant data, achieving precision with an approximation of 95.87% of the genealogical links.

Keywords; *Ancestry, descendant, genealogical linkage, machine learning, Bayesian networks.*

I. INTRODUCTION

Identify, family evolution through ancestry and descent over time, is a necessity for people because we are the result of many generations of inhabitants of the earth, who have transmitted life and experiences manifested in a growing population of the world that gives continuity to our days.

Genealogy allows us to identify religions, myths, and studies in different branches of science to search and find answers and faces clarification of the resulting information, which is achieved by reducing uncertainty. We all have a past and a present, which can be related through the genealogy that seeks to answer the question, where do we come from? (one of the three questions in Paul Gauguin's painting Where do we come from? Who are we? Where are we going?), considering the blood relationship that is the connection between people who descend from the same root, or are united by blood ties and that today also incorporates the DNA that keeps much information from our ancestors, for the transmission of hereditary characters where knowing our ancestors becomes a means to strengthen identity, memory, discover hereditary diseases, personality, and even discovers history; important for security, science, and history, even more so for the individual and society.

The basis for doing this research work is archival documents, and records as the testimony of the past, the proposed solution requires the construction of an intelligent genealogical linkage model in which various machine learning algorithms are applied to process the data which come from civil and identification registers, which is often interpreted by users as a learning model, which is a black box due to the functions and mechanisms that are sometimes incomprehensible and unclear, about why and what how the model works, for the development of the model with high-performance characteristics, it requires a large amount of processing time with a

Published by: The Mattingley Publishing Co., Inc.

May – June 2020
ISSN: 0193-4120 Page No. 17937 - 17946

trial and error approach, which becomes a great challenge to make it transparent, making it explainable and understandable, especially in the functioning of its internal mechanisms.

II. BACKGROUND

Family history is more than just dates of birth, marriage, and death; it is the history of the interaction of individuals with the community and others [Central Library Resource Council of New York 1997]. The emperors and kings of civilizations so distant and separated in space and time used genealogy to validate their claims to the throne, to trace their lineages back to the gods.

[1] in his thesis "From generation to generation: family stories, computers, and genealogy", it is proposed as a research objective to provide the theoretical framework to understand the importance and characteristics of family storytelling, its benefits, and its relationship with construction community. The procedures and techniques used to obtain stories, and the relevance of using them as keys to indexing and exploring existing systems, as well as analyzing the theoretical background to understand how an online community can build a story. In Peru, there are few attempts to carry out genealogy, with efforts that are not yet sufficient. The collection of information has its origin in the oral testimonies of members of the paternal family that have been gradually sought to contrast with the documentation in the civil registries and other historical sources that corroborate this information. As mentioned [2] in his interesting work about the absence in Peru of a central record of foreigners, during the last century and even until the middle of the 20th century, it is difficult to obtain a clear image of the subject.

In their work [3] "Computationally inferred long-term discovered genealogical networks with trends in mating" explain that genealogical networks or pedigree populations are generally studied by genealogists who want to know about their ancestors, but also provide valuable resources for disciplines such as digital demography, genetics, and computational social science. These networks are generally built manually through a time-consuming process, and that requires manual comparison with a large number of historical records; for this, he developed a computational method to infer large-scale genealogy networks automatically. The study that spanned 150 years shows the tendency of people to spouse a socioeconomic level selective mating.

[4] In his research on “Family Trees on the Web: A Search Engine User Perspective” he presents an extensive study on the evolution of textual content on the Web, showing how some pages are created while others are created using existing material, that shows that a significant fraction of the Web is a product of the latter case. The concept of a geographic Web tree, which each page of a web snapshot is classified into a component.

[5] in his book "Reconstruction of the population", he explains that people form societies and compose family ties and networks with social, economic, religious dimensions and coexist in homes and communities; people are born, marry, have children and die, they change direction and also develop careers, for study people are at the center of the problem and must know each other in the context of their complex relationships. The book aims to identify people and reconstruct populations in conditions where information is scarce, ambiguous, confusing, and sometimes erroneous, involving an effort by historians, social scientists, and linguists with in-depth knowledge of the complexity of the past.

[6] In his work "Study of techniques to recognize localization patterns of words, images, and documents", he explains that the vast collections of documents available in image format must be indexed for information retrieval purposes. Word localization is an alternative solution to Optical Character Recognition (OCR), which is inefficient at recognizing text of degraded quality and unknown

Published by: The Mattingley Publishing Co., Inc.

May – June 2020
ISSN: 0193-4120 Page No. 17937 - 17946

fonts that appear in printed text or variations of writing style in written documents.

[7] In the thesis "Physiology and methodical pluralism. A study on the genealogy of morals "addresses the relationship between physiology and method in Nietzsche's philosophy. Its purpose is to clarify how to proceed by tracing the origin of moral valuations in The Genealogy of Morality. By focusing on the connection between life and the value, The importance of the role played by the Nietzschean conception is highlighted. It explains the methodical pluralism used in genealogical work. Finally, it shows how the tools used in lines of the genealogy of morality.

[8] Explains that, depending on the scenario in which a problem develops, there can be many ways to solve it, since the approach helps us to clarify the path to the correct solution, in which we will find many algorithms to develop, which It will help to find not a correct answer, but the best possible solution, addressing the problem, according to the chosen strategy. Rational AI agents approach problems similarly, searching for the solution space to provide the best result.

2.1 Citizen records

The Birth Registry is a registry in charge of RENIEC, in which the birth of all Peruvians is registered, in addition to other acts, such as recognition of parents and adoptions, among others. The registry remains in the custody of RENIEC. The CNV birth certificate is a legal instrument that certifies the name of a person, besides, the Peruvian nationality and accredits filiation links. Parents individually or jointly can carry out this registration.

The national identity document (DNI) is a public, personal and non-transferable document, which constitutes the only title of voting right for the person in whose favor it has been granted; It is mandatory for all nationals. The DNI is granted to all Peruvians born inside or outside the territory of the republic from the date of their birth and to those who are nationalized since the nationalization process is approved. The document issued must assign a unique CUI identification code, which will remain unchanged until the person's death, as the only identification reference for the person.

III. METHODOLOGY

According to [9], the scientific method as it advances shows us the tendencies towards the future that challenges our ability to solve problems that will modify the fundamental structures on which the development of humanity rests (Pineda, 1996).

According to [10], basic research is a scientific and technological activity that discovers general laws, and that constitutes a type of research, within the context of scientific research that is related to applied research and experimental development. Primary research is based on the discovery that leads us to the finding that can explain a phenomenon, a research problem, a social or natural requirement.

Being a model, knowledge is added by adjusting perfectly to the particular characteristics of the typology, and this research also has application because it aims to offer innovative answers through theoretical concepts and apply them so that it can contribute to solving social problems with futuristic perspectives, through an integrated model of good business practices for organizations attached to the Lima stock exchange. Furthermore, according to Pino (2011), "applied research is also called practical or empirical, it is characterized by seeking the application or use of the knowledge acquired" (p. 253). For his part, [11] points out that "applied research seeks to know to do, act, build and modify, it is concerned about the immediate application to a specific reality" (p. 39).

3.1 Research level

The research level is descriptive, which seeks to select a series of questions and measure each one independently, to describe what is being investigated. According to [12], the researcher's objective is to describe phenomena, situations,

Published by: The Mattingley Publishing Co., Inc.

May – June 2020
ISSN: 0193-4120 Page No. 17937 - 17946

contexts, and events, detailing how they are and how they manifest themselves. Also, according to the nature of the study, it becomes explanatory research, since they are structured and determine the causes of the phenomena, generating a sense of understanding. [12] affirms that: Explanatory studies respond to the causes of events and physical or social events, explaining a phenomenon, and the variables are related.

The research method is framed according to the typology of analytical research since the different topics will be widely analyzed through their corresponding variables; Also, it applies the deductive method that will allow inferring valid ideas, concepts, theories and conclusions, and then drawing conclusions using a logical deduction, comparing each other and other relevant statements (Popper, 1997).

3.2. Variables and indicators

Independent variable civil registries: Evidence and documentation of data and facts that concern events and actions related to the civil status of individuals or physical persons such as registration of births, marriages, and deaths. Civil registries have reliable information about citizens, useful for protection, social assistance, others.

Independent Variable - Identification Record: The identification record is linked to identity, which is the set of characteristics of a subject, citizen, or community, which characterize the individual or the group, concerning the which is built throughout life. However, the process is particularly active during adolescence. Identification consists of the assimilation of a property or attribute of another person, transforming oneself under a personality.

Dependent Variable - Genealogical Links: The expected result of the genealogical links then applying the artificial intelligence model will allow the author to evaluate the relationship of kinship and the type of descent; that is, relationships are recognized on both the father's and mother's side. The family is made up of all those people with whom a genealogical or blood ties can be established. Marriage is essential because it strengthens and builds alliances between relatives.

3.3. Hypothesis test strategy

For the research, a statistical method will be applied that allows determining the presence or absence of association of the variables: Independent (X) called Civil registries and identification record and the dependent (Y) named genealogical links, the variables submitted to the present investigation, therefore, the hypothesis testing strategy is based on a general model shown below:

The chi-square distribution is a non-parametric statistical test that is used for hypothesis contracting, according to [12] “non-parametric tests are statistical procedures that can be used to test hypotheses when it is not possible to set any assumption on population parameters or distributions ”In this sense, the X2 test consists of goodness-of-fit tests and tests of independence.

Therefore, for the hypothesis test according to the objective of the research will be applied the X2 of independence, because it is a tool that allows us to check the independence of categorical variables; and it facilitates the analysis of two factors to determine the existence or not of a relationship between the variables. To do this, cross-tabulation or contingency tables are applied and used.

3.4 Research techniques

Techniques: As data collection techniques, it has been developed under the support of Montenegro (2012) "the technique is the set of steps or operations that are executed in a certain order and with certain instruments, to obtain a result" (p . 12). Therefore, in the research, the data collection techniques, according to [12] cited by [11]:

Surveys. They are the questionnaires to measure knowledge levels and attitude scales. (p. 194)

Published by: The Mattingley Publishing Co., Inc.

TEST Engineering & Management

May – June 2020
ISSN: 0193-4120 Page No. 17937 - 17946

Instruments: The data collection instruments applied corresponding to the techniques indicated above, having the following instruments:

Survey Form: Which will be used to both the General, Administrative and Head of Social Responsibility Area representatives of the organizations attached to the BVL; having formulated an average of 9 closed-type questions, related to the study variables, in particular, related to The application of the business process management methodology - BPM; and Questionnaire of the level of integration and implementation of business practices according to [11].

The measurement instruments in the present investigation are observation, and the variables will be manipulated, the documentary work will be focused on the review of books, magazines and other documents that will be related to the investigation and the bibliographic sheets because they will be used to annotate the data during the research process, likewise, [12] states that the survey is "a set of questions regarding one or more variables to be measured".

Processing: To process and carry out the analysis of the results of this research, tools such as Microsoft Excel, SPSS are applied, which are computer tools that will allow the results obtained from the field study to be presented in a detailed and appropriate way. According to the author [11], "Descriptive statistics using frequency tables and bar diagrams" will be used.

The Table 1 of this research project shows the variables operationalized, which will allow defining how each of its characteristics will be observed and measured in the research, making a conceptual definition with nominal and descriptive clarifications that describe the essence of real attributes of the object appropriate to the practical requirements of the study.

Table 1: Operationalization of variables.

Variables	Dimensions	Indicators	Instruments
Civil registries	Marriage	Number of marriages of the person Date of marriage Place of marriage	Civil documentation Military documentation RENIEC database Data transactions
	Births	Date of birth Place of birth	
	Deaths	Date of death Place of death	
	Emancipations	State of autonomy	
Identification Record	CNV (Live Birth Certificate)	Live Birth Certificate (CNV) Unique Identification Code (CUI)	Document national identity.
Genealogical Links	Birth Certificate	Degree of relatives Relationship level	Church records Family tree Military registry
	Ancestors	Number. of children	
	Descendents	Degree of consanguinity	
	Consanguinity	Degree of affinity	

To obtain the values of the indicators, the following instruments have been considered that will allow testing the hypotheses after processing the data.

IV. PROCEDURES

To apply the data collection and analysis methods, it is essential to carry out the corresponding evaluations, these evaluations go beyond the magnitude of the effects to determine with whom, and in what way the data will be obtained, for this, the contributing factors are examined in advance. To succeed in analyzing and synthesizing data to answer specific evaluation questions, with the combination of empirical evidence. In this phase of evaluation planning, the analytical framework is specified. In essence, the methodology for analyzing data by examining patterns systematically and transparently that involves using appropriate numerical and textual analysis methods and triangulating multiple data sources and perspectives to maximize the reliability of the data process.

A genealogical network consisting of many citizens is built, which is supervised by an individual genealogist over a long period.where, individuals are first related to the birth records in the dataset. A citizen is considered compatible if a record with the same normalized first and last name and the same date of birth is found. Then, the parental matches of children where both citizens match the birth record are searched to obtain the links. The links are divided into a training set and a test set. This is done by calculating the connected components of the

Published by: The Mattingley Publishing Co., Inc.

TEST Engineering & Management

May – June 2020
ISSN: 0193-4120 Page No. 17937 - 17946

network, sorting the elements by size in descending order, and assigning the nodes into two categories, in turn, assigning the most significant element to the training set, then assigning the most significant component to the test set. , alternating and skipping a bucket if its target size has been reached. Compared to directly dividing people into the two segments, this approach ensures that no edges are running through the two parts that would thereby be lost. In the end, the test links between mother and child are obtained.

4.1. Planning of the evaluation

Before defining what data to collect and how to analyze it. Clarified the purpose of the evaluation, high-level evaluation questions are identified, ideally, with input from key stakeholders; Assessment questions are prescribed by a previously developed assessment system or assessment framework. Answers to evaluation questions — regardless of how they are arrived at — should ensure that the purpose of the evaluation is met. Having an agreed set of questions guides what data to collect, how to analyze the data, and how to report the evaluation findings. Right evaluation questions not only ask, "what were the results?" (descriptive questions), but also "how good were the results?" as explanatory variables in the causal chain.

V. ANALYSIS OF DATA

Planning for data collection begins by reviewing existing data, and it is especially important to check for baseline data for the selected indicators, as well as for the social, demographic, and other relevant characteristics of the study population. The evaluation design involves comparing the changes over time of different groups, and the baseline data will be used to determine the equivalence of the groups before the program starts or to "match" different groups (as in the case of the designs quasi-experimental).

Table 2 Evaluation matrix: Data collection with the Key evaluation questions (KEQ)

Key evaluation questions (KEQ)	Citizen survey	Key user interviews	Information records	Observation of algorithm execution
KEQ 1 What was the quality of execution?		✓	✓	✓
KEQ 2 To what extent were the objectives met?	✓	✓	✓	
KEQ 3 What other impacts occurred?	✓	✓		
KEQ 4 How could the algorithm be improved?		✓		✓

Table 3. Collection options (primary data) and collation (secondary data) Own Source

Option	What could it include?	Examples
Retrieval of existing documents and data	• Formal policy documents, plans and performance reports • Official statistics • Monitoring data • Records.	• Review program planning documents, meeting minutes, progress reports. • The political, socioeconomic, or health profile of the country or the specific place where the study was carried out.
Collection of data from individuals or groups	• Interviews with individuals, groups, directed discussion groups, projection techniques. • Questionnaires or surveys via email, websites, face-to-face, mobile data. • Specialized methods (voting, card classification, seasonal calendars, projection techniques, experiences).	• Interviews with key informants between representatives of relevant government departments, non-governmental organizations or the development community at large • Interviews with program directors, program implementers, and those responsible for routine program supervision. • Interviews, group discussions (such as discussion groups), or questionnaires with program participants.
Observation	• Structured or unstructured • From participants or non-participants • Participatory or non-participatory • Recorded through notes, photos or videos	• Observations of program activities and interactions with participants.
Physical measurement	• Biophysical measurements • Geographical information	• Weight of infants • Places with a high prevalence of HIV infection.

Table 4.Summary of sampling options with illustrative methods.

Group of sampling options	Some specific methods	Risk of introducing biases
Probabilistic: Use random or quasi-random methods to select the sample and then use statistical extrapolability to conclude that population.	• Simple random sampling • Stratified sampling • Multi-stage sampling • Sequential sampling	This group presents specific rules on the selection of the sampling frame, the size of the sample, and the management of the variation in the sample.
Intentional: Study cases with abundant information from a certain population to draw analytical conclusions about the population. Units are selected based on one or more predetermined characteristics, and the sample size may be as small as one.	• Confirm and reject • Critical case • Maximum variation sampling • Valor atípico •Outlier snowball sampling based on theories of a typical case	This group encourages transparency in case selection and triangulation and seeks to reject empirical evidence.
Convenience: These sampling options use individuals who are available or cases as they occur.	• Easily available • Volunteers	This type has the lowest reliability but requires less time, funds, and effort.

VI. RESULTS

To test the hypothesis, clean data was processed from the database of birth certificates of the

Published by: The Mattingley Publishing Co., Inc.

May – June 2020
ISSN: 0193-4120 Page No. 17937 - 17946

province of Tumbes from the period of 1900 to 2019, as Figure 5

Table 5 Attributes of the structured data model.

ATTRIBUTE	DEFINITION
NUM_ACTA_NAC	Birth certificate number
CUI	Unique identification code
FECHA_CREA_ACTA_NAC	Date of birth certificate creation
FECHA_DE_NACIM	Date of birth
FECHA_DE_INSCRIPCCION	Registration date
SEXO_NACIDO	Sex of the born
PRIMER_APELLIDO_NAC	Surname
SEGUNDO_APELLIDO_NAC	Second surname
NOMBRES_NACIDO	Names of the born
LUGAR_NACIMIENTO	Place of birth
LOCALIDAD	Town of birth
UBIGEO_LOCAL	Birth Placement
PRI_APELLIDO_MADRE	First surname mother
SEGU_APELLIDO_MADRE	Second surname mother
NOMBRES_MADRE	Mother's names
TIPO_DOC_MADRE	Mother document type
NUM_DOC_MADRE	Mother document number
PRI_APELLIDO_PADRE	First surname father
SEGU_APELLIDO_PADRE	Second last name father
NOMBRES_PADRE	Father's names
TIPO_DOC_PADRE	Parent document type
NUM_DOC_PADRE	Parent document number
PRI_APELLIDO_DECLANTE1	First declarant last name 1
SEGU_APELLIDO_DECLANTE1	Declarant second surname 1
NOMBRES_DECLANTE1	Declarant 1 Names
TIPO_DOC_DECLAR_1	Declarant 1 document type
NUM_DOC_DECLAR_1	Declarant document number 1
VINCULO_DECLARANTE1	Declarant Link 1
PRI_APELLIDO_DECLANTE2	First declarant last name 2
SEGU_APELLIDO_DECLANTE2	Second declarant last name 2
NOMBRES_DECLANTE2	Declarant 2 names
VINCULO_DECLARANTE2	Declarant link 2
TIPO_DOC_DECLAR_2	Declarant 2 document type
NUM_DOC_DECLAR_2	Declarant document number 1
ORIGEN_REGISTRO	Registry origin

Tabla 6 Data quality evaluation.

First Last name	Second Last Name	Names	Observatión
LÓPEZ	LEÓN	MARIA	Incomplete name
LÓPEZ	LEÓN	MARÍA DOMINGA	Okay
LÓPEZ	LEÓN	MARIA DOMINGA	Name without tilde
GARAY	RENTERIA	ROSA ANGÉLICA	Okay
GARAY	RENTERIA	ROSA ANGELICA	Name without tilde
PEÑA	GARCIA	ROSA JULIA	Okay
PEÑA	GARCIA	ANGEL	Okay
PEÑA	GARCIA	LUIS ALBERTO	Okay
PEÑA	GARCÍA	BIENVENIDO	Second Surname with Tilde
GARCÍA		ROSA	No Second Last Name

Table 6 shows the data quality assessment, some examples of incomplete, incorrect, inaccurate, and irrelevant data found in the DB.

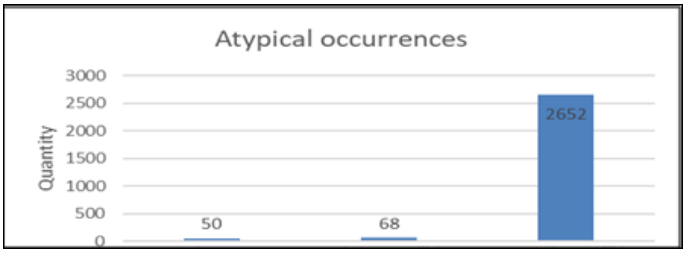

Figure 6.Consolidation of incomplete, incorrect, inaccurate, and irrelevant data.

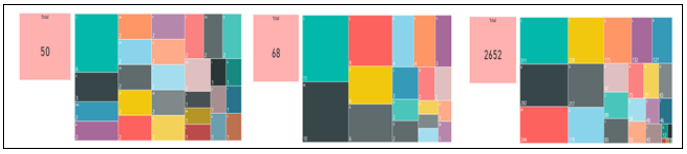

Figure 7 Visualization of incomplete, incorrect, inaccurate and irrelevant data

Figure 6 and 7 shows the number of atypical occurrences like incomplete, incorrect, inaccurate, and irrelevant data that has been identified, such as 50 names with one or two characters, 68 first last name with a single character, and 2 652 second last names with a unique character. These data are not considered when classifying them with Naive Bayes about incorrect fingering.

Table 7 Differences comparing surnames

First surname	Second surname	Name born	Difference Surname 1	Difference Surname 2	Difference Name born	Approx %
ROJAS	ARECHAGA	JULIO CESAR	1	1	1	1
ROJAS	ARECHAGA	JULIO CESAR	1	1	0.91	0.91
ROJAS	ARECHAGA	JULIO CESAR	1	0.88	1	0.88
ROJAS	ARECHAGA	JULIO CESAR	1	0.88	0.91	0.8008

Table 7 shows the differences when comparing paternal, maternal, and first names, where the value of 1 after processing the algorithms indicates no differences.

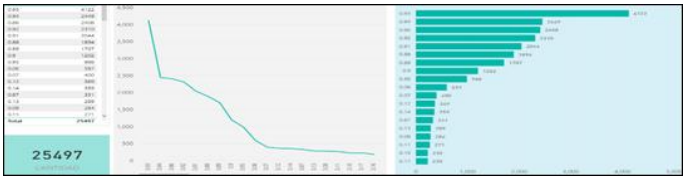

Figure 8 Detail of similarity in surnames and names

Published by: The Mattingley Publishing Co., Inc.

May – June 2020
ISSN: 0193-4120 Page No. 17937 - 17946

Figure 8 shows the detail of the inconsistencies when comparing paternal last name, maternal last name, and names of 25 497 processed records.

The testing of each hypothesis is carried out to implement an artificial intelligence model based on the identification and marital status record. Then it will verify the relationship and identity of the related parties; for this, the Naive Bayes classifier was implemented on the sample and managed to obtain the following results.

Table 8 Data evaluation.

Criterion Evaluation of the Data	Quantity	Percentage
Surname and name error characters	2 758	0.71%
Differences in sumames and names	25 497	6.59%
Correct data	358 425	92.69%
Analyzed data	386 680	100.00%

Table 8 shows the amount and percentage of data evaluated, where 92.69% is correct, and 7.3% is incorrect. Inconsistencies occur when comparing the last names and first names. Applied to the structured data, the precision is 92.69%, to improve the classification, the similarity in the names from 1% to 90% with the following results.

Tabla 9 Diferencias en Apellidos y Nombres.

Criterion Evaluation of the Data	Quantity	Percentage
Surname and name error characters	2 758	0.71%
Differences in sumames and names	13 201	3.41%
Correct data	370 721	95.87%
Analyzed data	386 680	100.00%

Table 8 shows the amount and percentage of data evaluated, where 92.69% is correct, and 7.3% is incorrect. Inconsistencies occur when comparing the last names and first names. When the is applied to the structured data, the precision is 92.69%, to improve the classification, those with a similarity in the names from 1% to 90% were taken into account, **obtaining the following results.**

Figure 9.Ingresos de datos de búsqueda.FuentePropia.

Detalle de la persona buscada, luego se evalua el detalle de la persona buscada, paraluegoidentificar a a la pareja, olasparejas, luego a los hijos, lasparejas de los hijos, los nietos y los bisnietos, el cua en estecaso no se tienendescendenciasyaque en esteúltimonivelaun no se tienenparejasvinculadas, como se muestra en la figura 10.

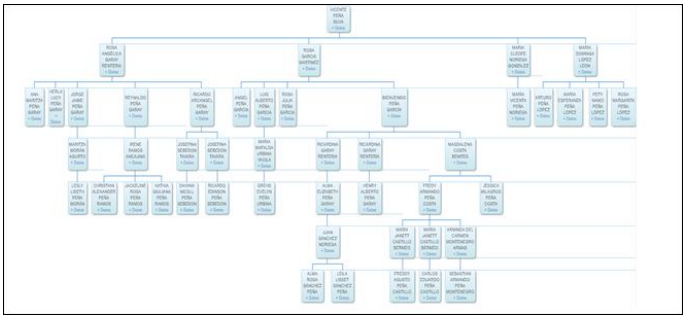

Figure 10.Niveles de identificacion de vinculosdescendientes

After observing the genealogy with the established relationships, the descending links of the citizen are determined, which reinforces the hypothesis.

6. 1 Analysis and data interpretation

Analyzed each of the tables and the previous figures, it is evident how the artificial intelligence model establishes the genealogical links of people based on the identification and marital status record. The effect is reflected in Figures 9 and 10, where the genealogical relationship is shown; this fulfills the objective and confirms the hypothesis.

6. 2 Discussion of results

To the internal validity, it is important to consider that the artificial intelligence model based on the identification and marital status registry has four stages: Data integration, data analysis, classification, and linking through Bayesian networks and genealogical visualization. Although it is true, all data from 1900 to 2019 from the department of Tumbes was taken as a sample

Published by: The Mattingley Publishing Co., Inc.

May – June 2020
ISSN: 0193-4120 Page No. 17937 - 17946

To external validity, the work of [3] on "Long-term genealogical networks discovered computationally inferred with selective mating trends" is considered as a reference, in which it achieves a link of 61.6% under the BinClass algorithm. This research has a link of 95.87%, due to having a structured data model and the directed learning of the algorithm that performs the respective links.

Take into consideration that information could not be collected from other data sources such as parishes, municipalities, among others, limiting the analysis to a structured data model, so the department of Tumbes was taken as a sample because of the data is greater completeness.

VII. CONCLUSIONS

When implementing an artificial intelligence model based on the identification and marital status registry, then, the relationship and identity of the related parties is verified from the year 1900 to 2019, for which the incomplete, incorrect, inaccurate and not relevant, found in the structured data model reaching, obtaining a 95.87% approximation in the links. The efforts to face the problems are related to the analysis and development of the intelligent model, which will show how genealogical visualization plays a critical role in its understanding, which is improved by applying machine learning techniques, which will allow users to understand and to value the efforts of Artificial Intelligence to help the citizen and increase knowledge.

VIII. FUTURE WORKS

Determine and design mechanisms to integrate structured and unstructured models that allow obtaining the relationship and identity of the related parties; in this way, it will be possible to have a holistic view of genealogical analysis.

IX. ACKNOWLEDGMENTS

Special acknowledgments to Eric Malmi for his collaboration and willingness to contribute information from his previous research, also to all the researchers of this project, especially Pedro Lezama and Danilo Chavez of RENIEC.

REFERENCES

[1] Hadis, M. (2002). From Generation to Generation: Family stories, computers, and genealogy. Massachusetts: Massachusetts Institute of Technology.

[2] Valverde Elera, J. (2015). Los Gallegos en el Perú. Universidade da Coruña: Cuadernos de estudios Gallegos, lxii.

[3] Malmi, E., Gionis, A., & Solin, A. (16 de Febrero de 2018). Computationally Inferred Genealogical Networks Uncover Long-Term Trends in Assortative Mating. Obtenido de arXiv® is a registered trademark of Cornell University: https://arxiv.org/pdf/1802.06055.pdf

[4] Baeza-Yates, R., Pereira, Á., & Ziviani, N. (1 de Enero de 2008). Genealogical Trees on the Web: A Search Engine User Perspective. Obtenido de Discover scientific knowledge and make your research visible: https://www.researchgate.net/publication/200110510.

[5] Bloothooft, G., Christen, P., Mandemakers, K., & Schraagen, M. (2015). Population Reconstruction. Netherlands: Springer.

[6] Giotis, A., Sfikas, G., Gatos, B., & Nikou, C. (2017). A survey of document image word spotting techniques. Greece: Department of Computer Science and Engineering, University of Ioannina.

[7] Cerna Solís, J. L. (2017). Fisiología y pluralismo metódico. Un Estudio sobre la genealogía de la moral. Lima: PUCP.

[8] Gupta, P. (10 de Enero de 2018). Hackernoon. Obtenido de Search Algorithms in Artificial Intelligence: https://hackernoon.com/search-algorithms-in-artificial-intelligence-8d32c12f6bea

May – June 2020
ISSN: 0193-4120 Page No. 17937 - 17946

[9] Pineda, E., de Alvarado, E. L., & de Canales, F. (1994). Metodología de la investigación. Manual para el desarrollo del personal de Salud. Washington: ORGANIZACIÓN PANAMERICANA DE LA SALUD.

[10] Carvajal, L. (2011). Metodología de la Investigación, Curso general y aplicado. Mexico: Poemia, su casa editorial; Edición: 28 (10 de agosto de 2011).

[11] Valderrama, S. (2016). Pasos para elaborar proyectos de investigación científica. Lima: San Marcos.

[12] Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, M. (2014). Metodologia de la investigacion. México D.F.: McGRAW-HILL/INTERAMERICANA EDITORES, S.A. DE C.V.

[13] J. Bustamante, C. Rodriguez, D. Esenarro. "Real-Time Facial Expression Recognition System Based on Deep Learning.". International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S11, September 2019

Published by: The Mattingley Publishing Co., Inc.