[1] **Alexander Fausto Medina Cortez**

[2] **Luis Alfredo Porras Tarifeño**

[3] **Pedro S. Castañeda**

Model for the Detection of Potential Car Insurance Customers by Applying Machine Learning Algorithms

Abstract: - The reduction in the production rate of the insurance sector in Peru is a problem that not only affects insurance companies, but also citizens, who cannot protect their valuable assets. This is due to the low acceptance of insurance in the Peruvian market and the inefficient use of the large amount of information and current technologies to detect clients by companies in this sector. That is why the need arises for a predictive model to identify which people are more likely to purchase insurance and what variables influence it. To do this, in this study, a solid model based on Machine Learning is proposed using data from a Peruvian insurance company, which is made up of sociodemographic variables and the vehicle that the person owns. The results show that the variables that influence being a potential insurance client are "employment status", "educational level", "salary range" and "vehicle size". Also, it was identified that the Decision Tree algorithm obtained the best performance, with an Area under the curve (AUC) of 0.88 and an F1-score of 0.82 in predicting potential clients.

Keywords: Automobiles, Potential customers, Machine learning, Insurance.

I. INTRODUCTION

The insurance sector plays a key role for citizens because it helps them protect their high-value assets from external risks that cannot be avoided such as theft, fire, accidents and even death, offering the security of obtaining financial compensation that reduces the impact of these events [1]. Because of this, there are products that protect people's vehicles, health, lives, and homes [2]. From a country's economic perspective, insurers help mobilize the domestic savings of its citizens, provide capital for external investments, facilitate trade and asset exchange, and contribute to the sustainable growth of a country [3].

Private insurance companies were recognized as the industry with the best and most sustainable growth until August 2021 in Peru, where they obtained a production index of over 400. However, the panorama is totally different since the post-pandemic period, reaching decrease to an approximate index of 200, almost 50% lower [4]. The decrease is due to different causes, including clients who have lost the perception of the importance of having insurance [5]. And on the part of the companies that have moved to a data-drive system, which is very primitive in Peru [6].

To solve this problem, various researchers have proposed solutions based on machine learning and data mining. For example, models based on unsupervised learning have been designed to segment customers and thus identify those with a greater probability of purchasing insurance [7], [8], [9], [10]. Furthermore, supervised learning techniques have been used to find the most robust prediction model [11], [12], [13]. However, these solutions have been proposed with data and context from other countries and between them they present different critical factors to define a potential client. Therefore, through this study, a predictive model applied to Machine Learning algorithms under the supervised learning paradigm is proposed in the Peruvian context.

This article is organized as follows: section 2 presents related works, where the existing literature is reviewed and compared. Section 3 presents details of the implementation methodology and validation results. Finally, section 4 presents the research conclusions and recommendations for future work.

II. RELATED WORKS

This section first considers the research that provides factors to identify potential customers. The factors can be divided into two categories: personal variables and vehicle variables (see Table 1).

The factors referring to "personal variables" are those that identify users such as "employment status" (the person has an independent, dependent or informal job) [7], [9], "salary range" (the person's salary determined by their socioeconomic level) [7], [9], [11] "gender" [8], [9], [11] "age" [8], [9], [10], [11], [12], [14] "geographical location" (place of residence of the person) [9], [10], [11], [14] and "study grade" (has a higher level of education) [7], [11], [12]. Regarding the "vehicle variables", the following were considered:

[1, 2, 3] Faculty of Information Systems Engineering, Universidad Peruana de Ciencias Aplicadas (UPC)

* Corresponding Author Email: [1]u201920348@upc.edu.pe, [2] u201920218@upc.edu.pe, [3] pcsipcas@upc.edu.pe

"bodywork" (vehicle structure) [8], [12], [14], "number of seats" (defines the size of the vehicle) [8], [9], [14], [15] and "use" (purpose of the vehicle).

Secondly, predictive algorithms for detecting potential clients are analyzed.

In the work [11] a comparison of Machine Learning algorithms on unbalanced data is carried out and the precision result of the following algorithms was obtained: logistic regression (LR) 84.42%, Gaussian Naive Bayes (GNB) 77.12%, random forest (RF) 82.69%, decision tree (DT) 76.54%, Support Vector Machine (SVM) 83.65%, k-nearest neighbors (KNN) 83.27%, Gradient boosting (GBM) 84.23% and XGBoost (XGB) 83.66%. On the other hand, in work [12], the following results were obtained for logistic regression accuracy (72%), decision tree (66%), RFC (70%), XGB (71%), SVC (74%) and MLP (76%). Furthermore, in the work [13], the following results were achieved for AUC decision tree (60.52), hybrid decision tree (81.17), random forest (54.80) and PS random forest (85.71).

Table I: Related works grouped by features

Feature	Related works
Employment status	[7] [9]
Salary range	[7] [9] [11]
Gender	[8] [9] [11]
Age	[8] [9] [10] [11] [12] [14]
Geographical location	[9] [10] [11] [14]
Study grade	[7] [11] [12]
Bodywork	[8] [12] [14]
Number of seats	[8] [9] [14] [15]

III. SYSTEM DESIGN

3.1 Architecture

In the auto insurance industry, the essential task of identifying potential customers has evolved over time. In the past, conventional strategies were often used, such as large-scale marketing campaigns with little personalization. Today, however, companies are leaning toward using artificial intelligence and machine learning to gain valuable data about consumers, allowing them to target their marketing strategies more precisely. Therefore, this article proposes an architecture that makes use of the information present in the company and uses the capabilities of the cloud to identify and show who is most likely to buy vehicle insurance.

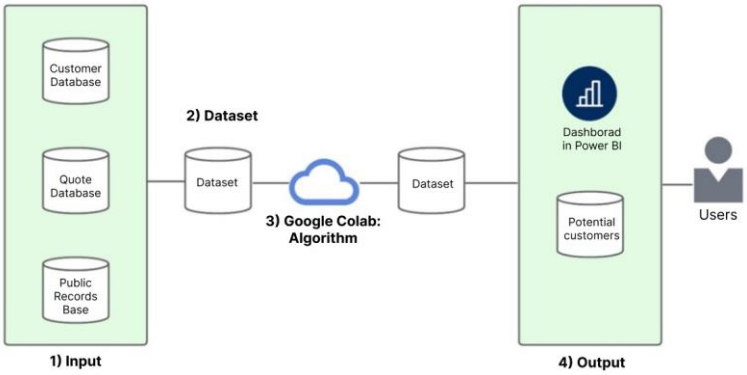

Fig. 1. Logical architecture of the machine learning model

1) Input

- Customer Database: This database contains information about current customers, including demographic data.
- Quote Database: Records of previous quotes are stored here, providing information on customer behavior in relation to their insurance.
- Public Records Base: This database includes public records information, such as driving records and vehicle ownership records, which are valuable for evaluating customer profiles.

2) Dataset

A consolidated data set was generated that unifies the information from the three sources mentioned above. Data consolidation plays an essential role in obtaining a comprehensive view of each prospective customer.

3) Google Colab: Algorithm

Google Colab provides a cloud platform for developing and running machine learning models. It offers access to scalable compute resources, which is essential for training sophisticated models on large datasets.

4) Output

The results of the model are presented through a dashboard in Power BI. This allows analysts and marketing teams to easily understand insights and make informed decisions. In addition, the model provides detailed information about potential customers, including their most relevant characteristics.

By integrating data from multiple sources, employing cloud resources and machine learning algorithms, insurance companies can significantly improve their ability to attract customers. The combination of Google

Colab and Power BI provides a comprehensive workflow for data-driven decision making and marketing strategy optimization.

3.2 Model

As part of the data collection from the three sources of income, the variables to be used in the model were collected. These variables are employment status, salary range, gender, age, study grade, seats, and vehicle use (see Table II).

Table II: List of features

ID	Feature	Description
VML01	Employment status	Type of Employment Contract
VML02	Salary range	The person's income
VML03	Gender	Gender of the person
VML04	Age	Years lived by the person
VML05	Study Grade	Maximum level of study
VML06	Number of seats	Number of usable seats in the vehicle
VML07	Vehicle Use	Purpose of the vehicle

These variables will be trained with the machine learning algorithms that obtained greater precision in other studies (see Table III): Random Forest (RF), Logistic Regression (LR) and Decision Tree (DT).

3.3 Indicators

In the quest to identify potential auto insurance customers effectively, machine learning models play a critical role. However, it is not enough to build a model; It is essential to validate its performance to ensure that decisions based on it are accurate and reliable. This section describes in detail the indicators and validation methods of the machine learning model designed for the auto insurance industry.

Accuracy, recall, and F1-score metrics were used to evaluate model performance. These metrics provided insight into the model's ability to correctly predict leads.

Table II: Machine learning algorithms

Algorithm	Description
Random Forest (RF)	It is a supervised learning algorithm that is based on the construction of multiple decision trees. Each tree is trained on a subsample of data and votes to predict the final label.
Logistic Regression (LR)	It is a supervised learning algorithm used primarily in binary classification problems.
Decision Tree (DT)	It is an algorithm that is used in both classification and regression problems. It works by dividing a dataset into smaller subsets based on key characteristics.

IV. EXPERIMENTS AND DISCUSSION

In this phase, we implement a rigorous performance evaluation. We use metrics such as precision, accuracy, specificity, and F1 Score to measure the model's ability to distinguish between potential customers and non-leads. In addition, we apply cross-validation to ensure the robustness of the model across different datasets, and we fine-tune hyperparameters to optimize their performance. The following table shows how the algorithms compare to the previously defined metrics (see Table IV).

It is observed that the algorithm that achieved the highest accuracy rate in the identification of potential car insurance customers was the Decision Tree (DT), reaching 82% in F1-score the same score as the Random Forest (RF), followed by logistic regression with 71%. In terms of specificity or recall, the most outstanding results were obtained with the Decision Tree (DT) and Random Forest (RF) algorithms, with rates of 90% and 89%, respectively. When it comes to model accuracy, all algorithms managed to surpass 76% in terms of predictive capability. In summary, the decision tree algorithm stood out as the leading algorithm in terms of accuracy and predictive capability across all the metrics evaluated.

Table II: Comparison of metrics by algorithm.

Alg	**AUC**	**Result**	**Precision**	**Recall**	**F1-Score**
LR	0.87	Potencial	0.74	0.86	0.80
		Non Potencial	0.79	0.65	0.71
DT	0.88	Potencial	0.90	0.75	0.82
		Non Potencial	0.75	0.90	0.82
RF	0.88	Potencial	0.89	0.75	0.82
		Non Potencial	0.75	0.89	0.82

V. Conclusions And Future Projects

A model's ability to detect potential auto insurance customers helps policyholders understand their target audience and better segment their market segment. In this article we propose a predictive analytical model based on machine learning to understand the potential of automobile insurance in Peru. We use the Random Forest, Decision Tree and Logistic Regression algorithms.

The dataset was obtained from a Peruvian auto insurance company. Likewise, an investigation and analysis of various variables that influence the purchase of automobile insurance was carried out. We observe that the probability of increasing is if the person's employment situation is dependent, has higher education, their salary range is A, B or C and if the vehicle has five or fewer seats.

To measure the performance of the model, evaluation measures were used: recall, precision, and F1-score. The algorithm that showed the best result was the decision tree, achieving a score of 82%.

As a future work, it is recommended to experiment with the model in different types of insurance such as life, health or home insurance. More algorithms can even be added to compare the level of accuracy with the three implemented in this article.

VI. Acknowledgements

The authors are grateful to the Dirección de Investigación de la Universidad Peruana de Ciencias Aplicadas for the support provided for the realization of this research work through the economic incentive

References

[1] Y. Grize, W. Fischer, and C. Lützelschwab, "Machine learning applications in nonlife insurance," Applied Stochastic Models in Business and Industry, 2020, pp. 1-15, doi:10.1002/asmb.2543.

[2] N. Dhieb, H. Ghazzai, H. Besbes, and M. Yehia, "A Secure AI-Driven Architecture for Automated Insurance Systems: Fraud Detection and Risk Measurement," IEEE Access, vol. 8, April 2020, pp. 58546-58558. doi:10.1109/ACCESS.2020.2983300.

[3] W. Teng Chang and K. Huong Lai, "A Neural Network-Based Approach in Predicting Consumers' Intentions of Purchasing Insurance Policies," Acta Informatica Pragensia, vol. 19, núm. 2, 2021, pp. 138-154. doi:10.18267/j.aip.152.

[4] National Institute of Statistics and Informatics (INEI), "Informe técnico de Producción Nacional," Lima: Instituto Nacional de Estadística e Informática (INEI), 2023. [Online]. Available at: https://bit.ly/41tCMVX.

[5] C. Bendayan Gamarra, N. Carhuaz Alarcon, P. Perez del Solar Zegarra, N. Peyre Alva, and R. E. Manchego Rosado, "Research and Proposal for an Advertising Campaign for Insurance," Bachelor's Thesis, Pontificia Universidad Católica del Perú, Lima, Peru, 2021. [Online]. Available at: https://bit.ly/41yS8cj.

[6] R. Villanueva, "Data, a common challenge for the insurance sector and its impact on the customer," Gestión, July 26, 2022. [Online]. Available at: https://bit.ly/40sOaAM.

[7] W. Qadadeh and S. Abdallah, "Customers Segmentation in the Insurance Company (TIC) Dataset," Procedia Computer Science, 19 de abril de 2018, pp. 277–290, doi:10.1016/j.procs.2018.10.529.

[8] F. Khamesian, F. Khanizadeh, and A. Bahiraie, "Customer Segmentation for Life Insurance in Iran," International Journal of Nonlinear Analysis and Applications, 2021, pp. 633-642, doi:10.22075/IJNAA.2021.22324.2350.

[9] S. Abdul-Rahman, N. F. Kamal Arifin, M. Hanafiah, and S. Mutalib, "Customer Segmentation and Profiling for Life Insurance using K-Modes Clustering and Decision Tree Classifier," International Journal of Advanced Computer Science and Applications, vol. 12, núm. 9, 2021, pp. 434-444, doi:10.14569/IJACSA.2021.0120950.

[10] Y. Liu, K. Bo, Q. Yi, Z. Wang, Y. Sun, J. Xu, and X. Zhang, "Predict Health Insurance Purchase with Machine Learning Techniques," SURF, 2023, pp. 1-16, doi:10.2139/ssrn.4366801.

[11] N. Kemboi Yego, J. Kasozi, and J. Nkurunziza, "A Comparative Analysis of Machine Learning Models for the Prediction of Insurance Uptake in Kenya," Data, vol. 6, núm. 11, 2021, pp. 116-132, doi:10.3390/data6110116.

[12] M. Codruţa Mare, D. Manaţe, G. Mureşan, S. L. Dragoş, C. M. Dragoş, and A. Purcel, "Machine Learning Models for Predicting Romanian Farmers' Purchase of Crop Insurance," Mathematics, vol. 10, núm. 19, 2022, doi:10.3390/math10193625.

[13] M. Uddin, M. Faizan Ansari, M. Adil, R. Chakrabortty, and M. Ryan, "Modeling Vehicle Insurance Adoption by Automobile Owners: A Hybrid Random Forest Classifier Approach," Processes, vol. 11, núm. 2, doi:10.3390/pr11020629.

[14] S. Mau, I. Pletikosa, and J. Wagner, "Forecasting next likely purchase events of insurance customers: A case study on the value of data-rich multichannel environments," International Journal of Bank Marketing, vol. 36, May 2017, pp. 1125-1144, doi:10.1108/IJBM-11-2016-0180.

[15] A. Soroush, A. Bahreininejad, and J. Van Den Berg, "A hybrid customer prediction system based on multiple forward stepwise logistic regression model," Intelligent Data Analysis, vol. 16, 2012, pp. 265-278, doi: 10.3233/IDA-2012-0523.

[16] A. Al Mamun, M. Khalilur Rahman, U. Thevi Munikrishnan, and Y. Permarupan, "Predicting the Intention and Purchase of Health Insurance Among Malaysian Working Adults," SAGE Open, Dec. 2021, pp. 1-18, doi:10.1177/2158244021106137.

[17] M. Hanafy and R. Ming, "Machine Learning Approaches for Auto Insurance Big Data," Risks, 2021, pp. 1-23, doi:10.3390/risks9020042.

[18] K. Hussain and E. Prieto, "Big Data in the Finance and Insurance Sectors," Cham, Switzerland: New Horizons for a Data-Driven Economy, 2016, doi:10.1007/978-3-319-21569-3.

[19] J. López Belmonte, A. Segura-Robles, A. J. Moreno-Guerrero, and M. E. Parra Gonzáles, "Machine Learning and Big Data in the Impact Literature. A Bibliometric Review with Scientific Mapping in Web of Science," Novel Machine Learning Approaches for Intelligent Big Data, vol. 12, 2019, doi:10.3390/sym12040495.
[20] I. Lozano Girón, "Coverage: How Much Do Peruvians Allocate for the Payment of an Insurance Premium?," El Comercio, Feb. 2019, 2020. [Online]. Available at: https://bit.ly/3M5QaLs.
[21] Z. Moradpour and G. Jandaghi, "Segmentation of life insurance customers based on their profile using fuzzy clustering," International Letters of Social and Humanistic Sciences, vol. 61, Oct. 2015, pp. 17-24, doi:10.18052/www.scipress.com/ILSHS.61.17.
[22] J. Perch Nielsen, V. Asimit, and I. Kyriakou, "Machine Learning in Insurance," London: Risk, 2022, doi:10.3390/risks8020054.
[23] A. Taha, B. Cosgrave, and S. McKeever, "Using Feature Selection with Machine Learning for Generation of Insurance Insights," Applied Sciences (Switzerland), 2022, doi:10.3390/app12063209.
[24] A. Taha, B. Cosgrave, W. Rashwan, and S. McKeever, "Insurance Reserve Prediction: Opportunities and Challenges," 2021 International Conference on Computational Science and Computational Intelligence (CSCI), Dec. 2021, doi: 10.1109/CSCI54926.2021.00120.
[25] L. Tsvetkova, Y. Bugaev, T. Belousova, and O. Zhukova, "Factors affecting the performance of insurance companies in Russian federation," Montenegrin Journal of Economics, 2021, pp. 17-2018, doi:10.14254/1800-5845/2021.17-1.16.
[26] N. Yego, J. Kasozi, and J. Nkurunziza, "A comparative analysis of machine learning models for the prediction of insurance uptake in kenya," Data, 2021, doi:10.3390/data6110116.
[27] J. Zhou, Y. Guo, Y. Ye, and J. Jiang, "Multi-Label Entropy-Based Feature Selection with Applications to Insurance Purchase Prediction," 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), June 2022, doi:10.1109/ICAICA50127.2020.9181921.