

# Human resource optimization using linear regression machine learning model: case study SUNAT

Salazar Marín Gloria<sup>1</sup>, Condori Obregon Patricia<sup>1,2</sup>, Palomino Vidal Carlos<sup>3</sup>

<sup>1</sup>Image Processing Research Laboratory (INTI-Lab) Universidad de Ciencias y Humanidades, Lima, Perú

<sup>2</sup>Business on Engineering and Technology, S.A.C. (BE Tech), Lima, Perú

<sup>3</sup>Facultad de Ingeniería, Universidad Tecnológica del Perú, Lima, Perú

---

## Article Info

### Article history:

Received Nov 3, 2022

Revised Mar 10, 2023

Accepted Mar 12, 2023

### Keywords:

Human resource allocation

Linear regression

Logistic regression

Machine learning

Support vector machine

---

## ABSTRACT

The continue searching for organization's process improvement for reduce cost and increase efficiency is a big challenge for organizations nowadays. This paper is about to recognize the importance of process improvement focusing in the right human resource allocation. The research predict best optime human resource allocation in the Superintendencia Nacional de Aduanas (SUNAT) in the chemical materials control area using a linear regression machine learning algorithm. This model was validated with recollected data in the SUNAT's control locations, the results were compared with historical data to determine their efficiency obtained a mean square error 0.434 that is lower comparing to logistic regression and support vector machine algorithm. This research recommend the implementation of this model in all SUNAT's controls locations in Perú.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



---

## Corresponding Author:

Palomino Vidal Carlos

Facultad de Ingeniería, Universidad Tecnológica del Perú

Av. Petit Thouars 116, Lima 15046, Perú

Email: carlospalomino@outlook.com

---

## 1. INTRODUCTION

Business organization has a continuous effort in improve their processes for reduce cost and resources to increase their revenue. One of the most important factors is optimize the human resource and just have the right quantity of workers in each processes or activity. This study is focused in determine the optimal human resource allocation in the chemical materials control area in Superintendencia Nacional de Aduanas (SUNAT). In the literature review is available many papers related to the importance of human resource allocation, many factors are consider for these aim like, quantity of workers or find a mix of skills that cover the job and the process's requirements [1]-[5]. Also in this literature shows the most using method to choose the optimal resource allocation are statistical methods and machine learning (ML) algorithm [6]-[10]. Some of the most algorithm used to predict human resource allocation are neural network algorithm used for predict human resource allocation for container terminal [11]-[13]. Other algorithm used are the decision tree and linear regression both combined in a multilayer perception algorithm were used for predict human resource allocation for different business processes [14]. The resource allocation permits organizations achieve organizational goals to improve cost, time or quality [7].

One of the most important problems in the chemical materials control area in SUNAT is determine the right quantity of workers required for the three activities: visual inspection, document inspection and inside inspection. This activities are needed to determine if the vehicles bring some of the prohibited materials inside, this prohibited materials are use for drug elaboration, when the traffic of vehicles is high a large row of cars

are accumulate, cause discomfort among drivers in this case the workers are not enough and sometimes let the vehicles pass without the correct supervision, but some days there is not so much vehicle traffic and workers are free of their labors. That is the reason for optimize the allocation of workers taking into account historical data of the interventions registers. The proposed solution is to implement an ML algorithm (linear regression), evaluating the data and determine the correct model for optimize the numbers of workers required for each weekday in the controls points. This investigation has a great relevance in SUNAT and in the public institutions of Perú not only for optimize the resources allocations also to demonstrate the implementation of data solutions are important to automate and improve proceses in public institutions and bring a better service to the society.

This investigation required data recollection and the parameters required in this data for the linear regression machine learning model are the day of intervention, the name of the day, the number of interventions realized, the number of workers assigned, the number of police officer assigned and the quantity of material confiscated in kilos. the measured data for this work is obtained in the control office of Asia (south of Lima) from January to June 2022. The number of row recollected were fifty thousand records that represent the number of interventions realizes during these period.

For the forecasting and the model implementation the data were split in two, 70 percent for training data and 30 percent for test data [15], the training data was used to define the model and determine the coefficients and variables for the linear regression algorithm, after the model is training the test data was used to evaluate the model and compare results, some metrics were calculates like R-squared (R2) and mean squared error (MSE) and compare with the results of others models. The tool that help with the data analysis was python with the libraries numpy, sklearn, as the two more important libraries for this process, the integrated development environment (IDE) was visual studio code. The methodology used for the development of research is detailed below.

## 2. METHOD

### 2.1. Linear regression algorithm

In this research, linear regression algorithm is used to predict the quantity of workers required to accomplish the supervision of tasks in chemical materials control offices in SUNAT. The regression analysis is a technique for modeling the relationship between variables used for engineering, physical and chemical issues and others. the regression analysis si the most widely used technique for statistical and predictions. The application of this model include: data description, parameter estimation, prediction and estimation and control [16]. The linear regression consist in analyze dependent and independent variables and the relationship between them that can represent as a line that cross the recollected data. The simplest linear regression model is the determine by the mathematical expression  $Y = \alpha + \beta X + \epsilon$  where  $\alpha, \beta$  is the regression coefficient,  $x$  is the independent variable and  $\epsilon$  the dependent variable [17]. Figure 1 shows the full pipeline for the prediction of the quantity of workers required to realize the activities in the the chemical materials control points in SUNAT.

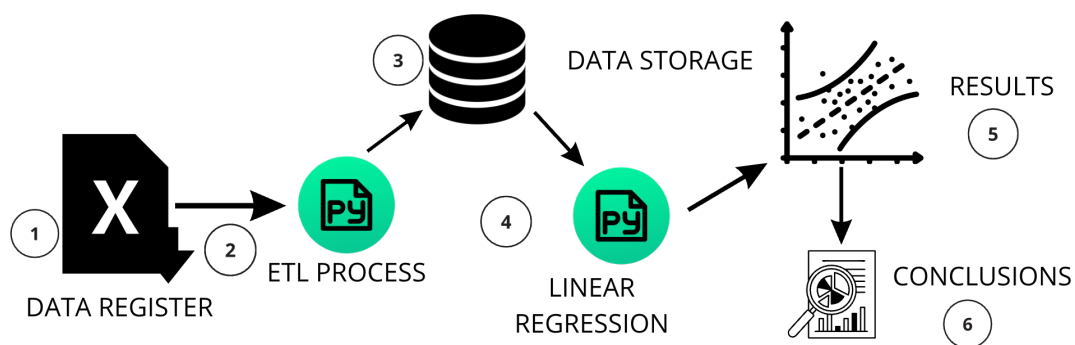


Figure 1. Pipeline for the linear regression prediction

Step 1: SUNAT has 10 checkpoints places distribute in Lima, Cusco, and Vraem. The information used for this analysis were registered on Asia checkpoints only. For collect the data an Excel sheet with visual

basic interface were used. The detailed information is on Table 1, the data contain fifty thousand rows from January to June 2022.

Table 1. Parameters

Variable	Description
Intervention date	The day of the intervention was performed
Name of week	the name of the seven days week
Number of interventions	The quantity of interventions per day
Number workers	Number of workers assigns that day
Number police officers	Number of police officers assigns that day
Material confiscated	The kilos of confiscated material found that day

Step 2: extract, transform and load process (ETL), this process is useful for prepare the data before the use in the analysis. The first phase is extraction get the clean data for different origins, the second phase is transformation the validation is performed to correct data errors and the last phase is the load, the new clean and validated data in a integrated repository [18]. For the implementation and use of the ETL process python was used and the data analysis was made using Python that is a free tool commonly used for data analysis and predictions [19].

Step 3: data base repository, the obtained data for the ETL process needs to be storage in a repository or database. The database selected for this analysis was SQLite, this tool is the less complicated transactional database to used in projects but keep the most important benefits of others transactional database like durability and faster performing [20]. The data was split and storage in three tables, one for the chemical materials allowed and don't during the inspections, another one for register the data of human resource participation in inspections and the last one to storage the detailed information of the inspections.

Step 4: data analysis and algorithm selection, in this step the data is slip in train data and test data, for determinate the correct coefficients of the linear regression algorithm. First the relationship between the independent variables and the dependent variable were analyzed, this case is a supervised learning algorithm [21], [22] because the dependent variable to predict is identified (number of workers), for determine the relationship with others variables. The variables name of week and numbers of interventions are the ones who had the higher correlation (+-.77) [23]. After determine the variables needed for the model a sklearn library [24] was used in python to train the model with this variables as a result the following mathematical model were obtained (1).

$$\text{Numberofworkers} = -0.61\text{Numberofday} + 0.21\text{Numberofinterventions} + 6.89 \quad (1)$$

This result were made using the train data, for validate the results the test data were used comparing the calculate results with the real results, and MSE metric were calculated [25]. But also the logistic regression algorithm [26] and the support vector machine algorithm (SVM) [27] were evaluated to compare which algorithm is more accuracy to predict the number of workers. The linear regression algorithm has the lowers errors (Table 2).

Table 2. MSE comparison

Algorithm	MSE
Linear regression	0.434
Logistic regression	0.5.43
Support vector machine	0.65

Steps 5 and 6: in this two finals steps the final reports are elaborated and presented to the organization. The results included predictions for the next three months with the optimized number of human resource required for the Asia checkpoint. Also a report with the detailed steps to update the data and de model to improve the model.

In (1) the full mathematical equation is shown. This equation determines the forecast values (number of workers). The representation of this equation is on Figure 2. In Figure 2 the line between the data is drawn and show the regression for the predicted values.

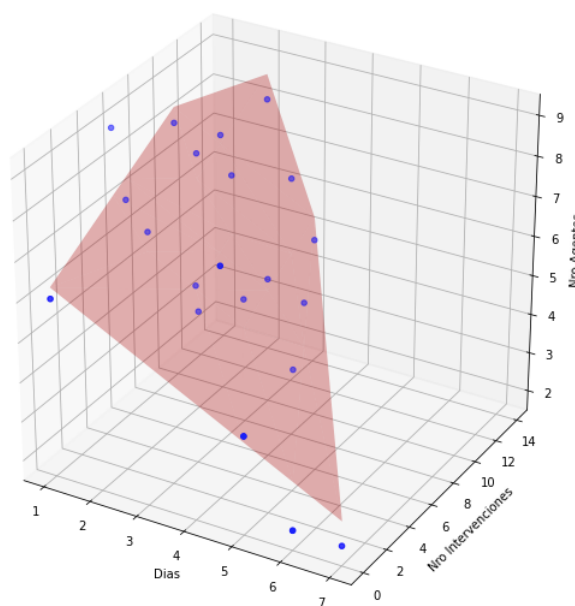


Figure 2. Graph of linear regression

### 3. RESULTS AND DISCUSSION

#### 3.1. Results of the prediction

As a final step with the linear regression model validated new data is evaluated with the model to predict the optimized number of workers required per day of the week. The results are shown in Table 3. The first days of the week required more quantity of human resource on average these days seven persons are required and during the weekends the number of human resource is decrease in average 4 persons are required.

Table 3. Result of the linear regression

Day	Number of workers
Monday	7
Tuesday	8
Wednesday	6
Thursday	6
Friday	5
Saturday	4
Sunday	3

#### 3.2. Discussion

The data in Table 3 indicate that during the weekends there is not much traffic of vehicles to control, therefore, so is possible reduce the workers required or assign them to other activities like a mobile control team to control vehicles in rotative locations. Due to the constant temperature changes during the weather seasons in Perú, vehicle traffic could change, so it is recommended to recalculate the algorithm every three months, which is the number of months that each weather season lasts. To obtain more benefits the six step of this analysis has to be replicated in others checkpoints places in Lima, Cusco, and Vraem and the recalculation time of the algorithm has to be according the weather of the checkpoints.




### 4. CONCLUSION

In this study the number of workers required for the activities of control were predicted. The data collected are from the last six months at the Asian checkpoint south of Lima. With this new human resource allocation was possible to optimize the performance of each worker by making them not only support in interventions but also in administrative and management work.




## REFERENCES

- [1] P. Ballesteros-Pérez, F. T. T. Thua, and D. Mora-Melià, "Human resource allocation to multiple projects based on members' expertise, group heterogeneity, and social cohesion," *Journal of Construction Engineering and Management*, vol. 145, no. 2, 2019, doi: 10.1061/(ASCE)CO.1943-7862.0001612.
- [2] J. Redpath and F. Nagia-Luddy, "'Unconscionable and irrational': SAPS human resource allocation," *South African Crime Quarterly*, no. 53, 2015, doi: 10.4314/sacq.v53i1.2.
- [3] M. Arias, R. Saavedra, M. R. Marques, J. Munoz-Gama, and M. Sepúlveda, "Human resource allocation in business process management and process mining," *Management Decision*, vol. 56, no. 2, pp. 376–405, 2018, doi: 10.1108/MD-05-2017-0476.
- [4] S. Mzakwe, "Equitable allocation of police human resources: social justice coalition and others v minister of police and others," *SA Crime Quarterly*, no. 69, 2020, doi: 10.17159/2413-3108/2020/vn69a7232.
- [5] Z. Wang, "Design of the human resource optimization allocation model based on information integration," *Mobile Information Systems*, vol. 2022, pp. 1–11, 2022, doi: 10.1155/2022/6549647.
- [6] K. P. Murphy, *Machine learning: a probabilistic perspective*. The MIT Press, 2012.
- [7] E. J. Kieling, F. C. Rodrigues, A. Filippetto, and J. Barbosa, "Smartalloc," in *Proceedings of the 25th Brazilian Symposium on Multimedia and the Web*, 2019, pp. 365–368, doi: 10.1145/3323503.3360643.
- [8] Y.-H. Xu, J.-W. Xie, Y.-G. Zhang, M. Hua, and W. Zhou, "Reinforcement learning (RL)-based energy efficient resource allocation for energy harvesting-powered wireless body area network," *Sensors*, vol. 20, no. 1, p. 44, 2019, doi: 10.3390/s20010044.
- [9] W. Shi and Q. Li, "Human resources balanced allocation method based on deep learning algorithm," *Scientific Programming*, vol. 2021, pp. 1–9, 2021, doi: 10.1155/2021/4681959.
- [10] S. Yuan, Q. Qi, E. Dai, and Y. Liang, "Human resource planning and configuration based on machine learning," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–6, 2022, doi: 10.1155/2022/3605722.
- [11] G. Fancello, C. Pani, M. Pisano, P. Serra, P. Zuddas, and P. Fadda, "Prediction of arrival times and human resources allocation for container terminal," *Maritime Economics & Logistics*, vol. 13, no. 2, pp. 142–173, 2011, doi: 10.1057/mel.2011.3.
- [12] Q. Feng, Z. Feng, and X. Su, "Design and simulation of human resource allocation model based on double-cycle neural network," *Computational Intelligence and Neuroscience*, vol. 2021, pp. 1–10, 2021, doi: 10.1155/2021/7149631.
- [13] R. Fang and Z. Fang, "Analysis of human resource allocation scheme for digital media big data based on recurrent neural network model," *Journal of Sensors*, vol. 2022, pp. 1–11, 2022, doi: 10.1155/2022/3430933.
- [14] W. Zhao, S. Pu, and D. Jiang, "A human resource allocation method for business processes using team faultlines," *Applied Intelligence*, vol. 50, no. 9, pp. 2887–2900, 2020, doi: 10.1007/s10489-020-01686-4.
- [15] C. M. Anderson-Cook, L. Lu, K. L. Myers, K. R. Quinlan, and N. Pawley, "Improved learning from data competitions through strategic design of training and test data sets," *Quality Engineering*, vol. 31, no. 4, pp. 564–580, 2019, doi: 10.1080/08982112.2019.1572186.
- [16] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*, 5th ed. Wiley, 2012.
- [17] Z. Yuan, "Interactive intelligent teaching and automatic composition scoring system based on linear regression machine learning algorithm," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 2, pp. 2069–2081, 2021, doi: 10.3233/JIFS-189208.
- [18] M. Souibgui, F. Atigui, S. Zammali, S. Cherfi, and S. B. Yahia, "Data quality in ETL process: a preliminary study," *Procedia Computer Science*, vol. 159, pp. 676–687, 2019, doi: 10.1016/j.procs.2019.09.223.
- [19] P. Virtanen *et al.*, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, 2020, doi: 10.1038/s41592-019-0686-2.
- [20] J.-H. Park, G. Oh, and S.-W. Lee, "SQL statement logging for making SQLite truly lite," *Proceedings of the VLDB Endowment*, vol. 11, no. 4, pp. 513–525, 2017, doi: 10.1145/3186728.3164146.
- [21] S. R. Kulkarni and B. Rajendran, "Spiking neural networks for handwritten digit recognition—supervised learning and network optimization," *Neural Networks*, vol. 103, pp. 118–127, 2018, doi: 10.1016/j.neunet.2018.03.019.
- [22] S. Zeng, Z. Liu, and X. Yang, "Supervised learning for parameterized Koopmans–Beckmann's graph matching," *Pattern Recognition Letters*, vol. 143, pp. 8–13, 2021, doi: 10.1016/j.patrec.2020.12.012.
- [23] R. S. Khedikar and A. S. Khobragade, "Extract texture features and correlation based on liner regression model using wavelet transform," in *Proceedings of the International Conference and Workshop on Emerging Trends in Technology*, 2010, pp. 1003–1004, doi: 10.1145/1741906.1742163.
- [24] W. A. van Eeden *et al.*, "Predicting the 9-year course of mood and anxiety disorders with automated machine learning: a comparison between auto-sklearn, naïve Bayes classifier, and traditional logistic regression," *Psychiatry Research*, vol. 299, p. 113823, 2021, doi: 10.1016/j.psychres.2021.113823.
- [25] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? – arguments against avoiding RMSE in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014, doi: 10.5194/gmd-7-1247-2014.
- [26] C. Mood, "Logistic regression: why we cannot do what we think we can do, and what we can do about it," *European Sociological Review*, vol. 26, no. 1, pp. 67–82, 2010, doi: 10.1093/esr/jcp006.
- [27] W. C. Leong, A. Bahadori, J. Zhang, and Z. Ahmad, "Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM)," *International Journal of River Basin Management*, vol. 19, no. 2, pp. 149–156, 2021, doi: 10.1080/15715124.2019.1628030.




**BIOGRAPHIES OF AUTHORS**

**Lic. Salazar Marín Gloria**    degree in mathematics from Federico Villarreal National University graduated in mathematics from the Federico Villarreal National University, a student of the Faculty of Law and Humanities of the Señor de Sipan University, extensive experience working at SUNAT checkpoints and chemical supplies in Peru, in the oversight area of the Ministry of Transport of Peru, as well as a social worker in schools and children's villages. She can be contacted at email: gloriasamar@gmail.com.



**Ing. Condori Obregon Patricia**    systems engineer at the University of Sciences and Humanities of Peru (UCH). CIO of the company business on engineering and technology, S.A.C. (BE Tech), expert in mobile development and machine learning projects, expert in planning, organization, development and constant feedback tasks aimed at the management of tools in the management and development of projects under the study of user behavior. She can be contacted at email: ccondori.patricia@gmail.com.



**Dr. Palomino Vidal Carlos**    doctor in administration (Federico Villarreal University). Master in direct and manage information technologies (Science Applied University). Graduated from the master's degree in information technologies (University Politecnica cataluña). Systems engineer (Federico Villarreal University). Project manager with 10 years experience with PMP 1643369 and PMI-ACP 2782702 credentials from PMI. Professor at Sciences and Humanity University, Callao University, Federico Villarreal University, Technological University of Peru. He can be contacted at email: carlospalomino@outlook.com.