Integrity of Training Data for Federal Civil Employees in Brazil

Flávio Garcia Praciano, Bruno J. G Praciano, Fábio L. L de Mendonça
Erica Lima Gallindo, Daniel Alves da Silva, Francisco C. M Duarte Jr. and Rafael T. de Sousa Jr
Department of Electrical Engineering, University of Brasília (UnB), Brasília-DF, Brazil
National School of Public Administration (Enap), Brasília, Brazil
Email: flavio.praciano@redes.unb.br

Abstract — **The Virtual School of Government: EV.g linked to the National School of Public Administration (ENAP) is an institution that aims to provide training for public servants in the three Government waits, as well as professional training for career positions. It aims to discuss the training amount of public servants and others. Qualitative bibliographic research was carried out for descriptive purposes, using an EmNumeros panel built with the Tableau tool and Python's data-mining program. For the progress of this article, they highlighted the importance of EV.g, acting as tools to enhance the civil service, by providing the qualification and knowledge of these professionals from the three spheres of government. have a fundamental role in the progress of the performance of public servants, which represents an evolution of services for the community and, as a result, increases the quality of services offered by the Public Administration. Consequently, it can be proven by the numbers informed in this article that the EV. g describes it as a means of reliably validating the training of civil servants, aiming at building active and responsible civil servants in their role in Public Administrations.**

Keywords - EmNumeros, Enap, Ev.g, Training, Public Servants, Enrollments, Schools of Government.

I. Introduction

In Brazil, there are approximately 11.4 million public employees. The total number increased from approximately 5.1 million to 11.4 million from 1986 to 2019. Divided among the three spheres of government (Union, States, Federal District, and Municipalities), in the three powers (Executive, Legislative, and Judiciary, being these three types: (public servants, public employees, and temporary employees). These data are available on the panel Em Numeros. In addition, a portal of courses offered by the Federal Government through the Virtual School of Government (EV.g) linked to the National School of Public Administration (ENAP) with various courses made available to civil servants of all expectations and communities. This availability is accurate data on the qualifications acquired by these publics [1].

In this context, the Federal Government, through EV.g, concentrates a set of information on the training courses held, which may raise public interest, such as thematic areas, courses held, demand and public profile for the training offered, agencies that use these training courses, number of employees trained, and others. Thus, when the efforts to make this information publicly available, social control is encouraged and the analysis of information according to the most varied needs. So far, this unified base stores about 4.3 million registrations made between 2006 and September 2021 [2].

EV.g has its educational management system, available online, in which interested parties identify existing training courses, register to attend them, and perform the associated activities to complete the courses. From this environment, students have access to their course conclusion certificates and can check the validity of the documents issued by EV.g [3].

Through such a portal, since 2017, EV.g has used the culture of active transparency about the service provided, making its information publicly available without the need for a prior request from interested individuals [2, 4].

EV.g has adopted the culture of transparency through its portal, which provides access to data on training offered, enabling interested parties to make their analysis of the information provided and fostering social control [4].

The availability of information about training held in the scope of EV.g meets the Access to Information Law (LAI) [4], which establishes as a duty of public agencies to promote the disclosure of information of public interest, regardless of requests, so that the culture of social control and transparency established in public agencies.

The remainder of this paper is organized as follows. Section II presents the related works. Section III describes the method for data analysis. In Section IV, the results are presented and discussed. Finally, Section V concludes the paper.

II. Related works

The authors of [5] performed a study of the data related to the information present in the communication channels of ENAP. Although the referenced work verifies the data of only one course, in this article, the proposal indicates all the possibilities of information from the data coming from the databases.

In Ferrarezi et al. [6], it was highlighted the creation of a system to register the history of courses taken by each public servant and a single register with possibilities to unify all the users' information in a single place. However, this paper created no system, but we created an Em Numeros panel for management visualization and a data mining code.

[7] shows the importance of training for public servants, but it does not show real numbers, and the objective of this article is data mining for an understanding visualization for managers for various increments of new actions, thus evidenced by a panel called Em Numeros.

The authors from [8] researched approximately 1923 public servants who did not complete one or more distance learning courses developed by ENAP, more significantly the evasion in distance learning courses, and one-third of them responded to the survey via email. A Statistical Package for Social Sciences (SPSS) [9] data collection tool was used for statistical analysis from the information acquired in the survey sent to the public servants, and the results were as follows: women (53.9%) dropped out more from distance learning courses, primarily from courses without tutoring (54.8%). In contrast, men drop out more from courses with tutoring (52.3%). In this paper, the proposal is very similar in which concerns informing data for decision making. Furthermore, the article itself reports mining of open data available for any information regarding the static of the courses and the student's information about the courses offered to them.

In [10], the authors did exploratory research through a Survey with help from ENAP. They analyzed several distortions of incentive for the Federal Civil Servants where it was found a wage discrepancy between careers with the executive power and the three powers with similar responsibilities. This paper cited and evidenced that there is wage inequality between careers, but the authors do not deal with some training or even a breakthrough in promoting course incentives for these professionals Federal Public, but this paper shows real numbers of how many employees are performing training and which spheres they are part.

[11] has a proposal for an automatic plugin for encouraging messages and optimism for the Moodle platform. This proposal aims to reduce the dropout rates of students in courses registered by them. The proposal also aims to reduce the rates of these dropouts in courses offered by EV.g. The proposal of this article follows. Differently, the cited presents a message plugin in Moodle, and this article analyzes and reports in real numbers after data mining processing from the Moodle platform of the National School of Public Administration (ENAP). These results are available in a panel called Em numeros: https://emnumeros.escolavirtual.gov.br/ for management decisions within the ENAP.

III. MATERIAL AND METHODS

This section details the methods for checking the data integrity and extends it to the machine learning approach. As shown in Figure 1, the analysis is divided into four functional blocks.

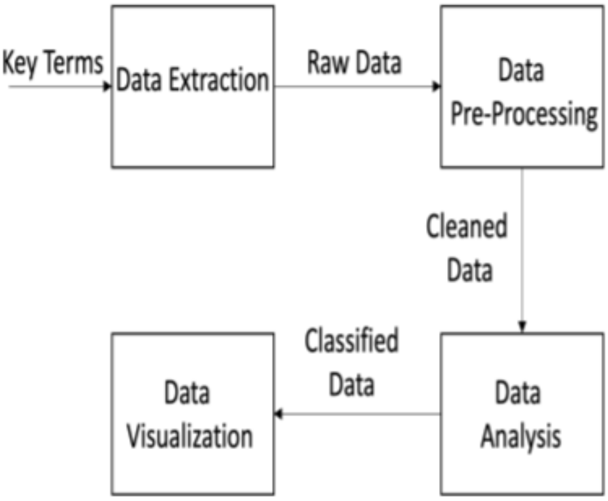


Figure 1: Block diagram of the proposed analysis for data integrity of Em Numeros Portal

A. Data Extraction

It is possible to visualize the cities of origin of those enrolled in the EV.G. courses gives territorial coverage in the Northeast, Southeast, and South regions, highlighting the large volume of enrolled students coming from the capital cities of these states. Still, in the indicators panel, as a first test of the public profile for the courses offered within the scope of the EV.g, we tried to identify the level of enrollment recurrence, what is the percentage of individuals who attend EV.g courses more than once.

The panel presents a set of filters that give flexibility to the user to perform its analyses, allowing to filter the graphics by year of enrollment (2006 to 2021), the sphere of government (federal, state, or municipal), the sphere of power (executive, judiciary or legislative), theme, course, and offer (class). The restructuring of the panels reorganized the information employing more dynamic visualizations of the enrollments and courses held within the scope of EV.g. In addition, based on feedback about the published information, the portal may provide improvements in the services provided, visually presenting information that may be of interest both to the EV.g management and to any ordinary citizen.

B. Data Pre-Processing

The data pre-processing used Pandas, where we removed the inconsistent data and the outliers present in some courses. It is essential to remember that the personal data presented in the dataset were removed. Regarding the outliers, we used the Boxplot approach, where we can see the median and interquartile ranges in a good way of visualization. Figure 2 shows an example of the outliers of the age of the enrolled students.

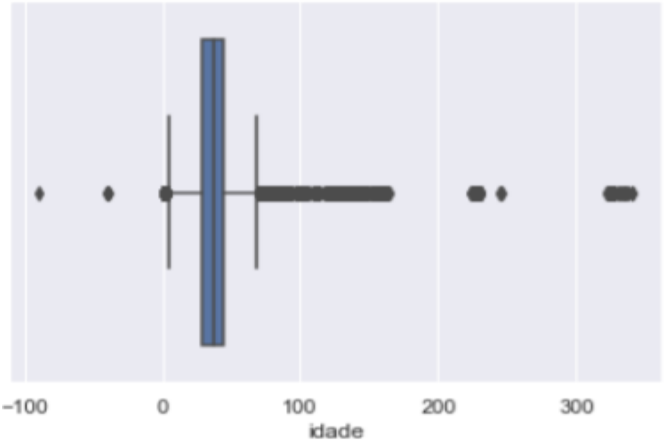


Figure 2: Boxplot with the data before the data preprocessing

We used the Z-Score metric for this outlier removal which removes the value of observation above the mean value [12]. It is necessary to perform it because we avoid overfitting in our estimator.

C. Data Analysis

We propose a tool that can predict the number of students that will finish the course. We use supervised machine learning algorithms once we have labeled data. Furthermore, we chose regression algorithms for this task. The analysis uses the Random Forest Algorithm [13], and to find the correct amount of estimators was computed with an array of [4, 8, 16, 32, 64, 128, 256].

In addition to the mentioned machine learning algorithm, we decided to compare Support Vector Machines (SVM) and an Artificial Neural Network (ANN) of two hidden layers. After that, we got the training results for the four classes we defined for the first interaction (Dropout, Complete, Fail, Locked, and Not Completed). For the second training step, we defined to merge the classes into just two approved, which contain the complete status, and another class not-approved with the dropout, fail, locked, and not completed label. Once the dataset contains more than 27 columns, it is necessary to understand how the main features are correlated.

The proposed ANN for this work was divided into four layers, where the input layer contains five perceptrons. Figure 3 shows the proposed ANN. In this case, since we have a binary classification (Approved/Failed), we may simply use a single-perceptron Output layer; If the output is smaller than 0.5, the student is approved; otherwise, the student fails.

For each layer, we have as an input a matrix made by columns of features (in our example, we have four features, i.e. Student Age, Course Duration, Student Gender, and Course Workload), that we label as I=1, 2, 3, 4. Each of this features will have n entries each feature is a vector $\{x_{(I)}\}_i$. The layer will have p perceptrons, labeled by a=1,…,p. Thus the output of the whole layer is a matrix $O_{(I)}^{(a)}$ given by:

$$O_{(I)}^{(a)}=\varphi(w_{(I)}^{(a)} \cdot x_{(I)}^{(a)}+b_{(I)}^{(a)}), \quad \text{(I)}$$

where w is the weight vector and b is the bias.

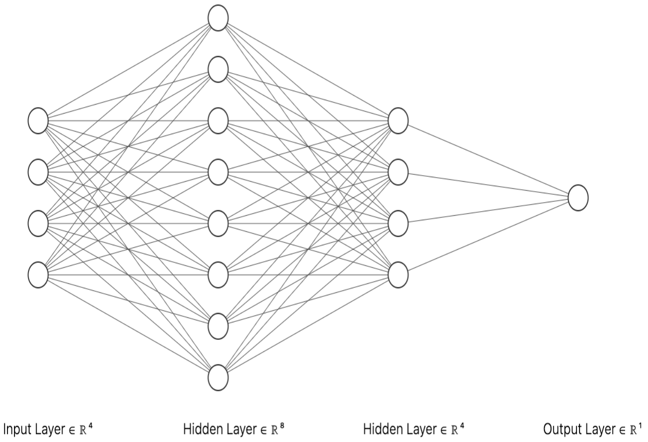
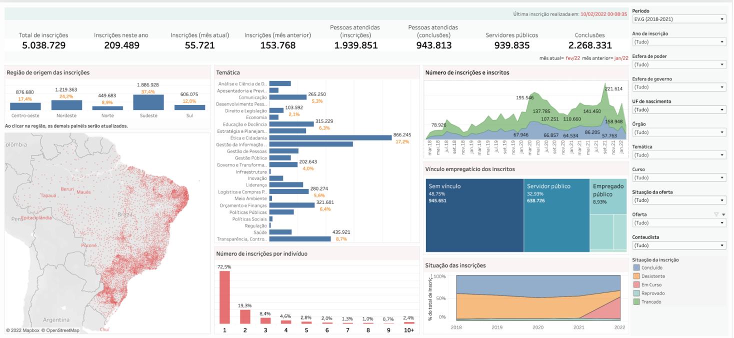

Input Layer $\in \mathbb{R}^4$
Hidden Layer $\in \mathbb{R}^8$
Hidden Layer $\in \mathbb{R}^4$
Output Layer $\in \mathbb{R}^1$

Figure 3: Proposed ANN architecture

D. Data Visualization

The data downloaded from Portal Em numeros were used to create some charts to support the responsible for each course, follow the metrics, and take action at the right moment. Since the mentioned portal provides visual analysis, we needed to create another kind of visual analysis and a way to understand the students’ behavior of some courses. Figure 4 shows a dashboard with the data available in this open database.


Última inscrição realizada em: 10/02/2022 00:08:35

Total de inscrições 5.038.729
Inscrições neste ano 209.489
Inscrições (mês atual) 55.721
Inscrições (mês anterior) 153.768
Pessoas atendidas (inscrições) 1.939.851
Pessoas atendidas (conclusões) 943.813
Servidores públicos 939.835
Conclusões 2.268.331

mês atual= fev/22 mês anterior= jan/22

Período: EV.G (2018-2021)
Ano de inscrição: (Tudo)
Esfera de poder: (Tudo)
Esfera de governo: (Tudo)
UF de nascimento: (Tudo)
Órgão: (Tudo)
Temática: (Tudo)
Curso: (Tudo)
Situação da oferta: (Tudo)
Oferta: (Tudo)
Conteudista: (Tudo)

Ao clicar na região, os demais painéis serão atualizados.

Figure 4: Dashboard available on Portal Em Numeros

IV. Results

As mentioned in Section III, we used three different algorithms for comparison. We could verify the consistency of the data of a specific course because we chose the following features from the dataset: gender, the course workload, age of the student, and duration of the course. These features were selected using a Principal Component Analysis, where it was possible to see the feature’s interactions. After that, it was possible to perform the training of the machine learning model using a different number of classifiers. The dataset was divided into training based on 70% for training, 20% for testing, and 10% for validation.

Figure 5 shows the average loss per epoch for the Neural Network.

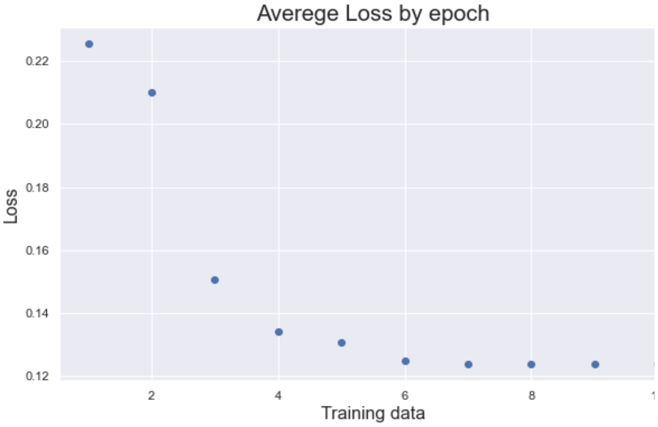


Figure 5: Average loss by epoch

We decided to use the precision metric of each algorithm for evaluation, because we just have two goal variables, such as Approved and Failed. This metric also gives us a measure of the relevant data points [14] [15]. It is important that we do not start treating a student who actually reproved as approved, but our model predicted as having it. Equation II defines this metric as a fraction of True Positives by the sum of True positives with False Positives.

$$Precision = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FP)} \qquad \text{(II)}$$

After the setup of our environment test, we could compute the selected metric for the three algorithms. Table I shows the performance of them. We compared an ensemble algorithm (Random Forest), linear algorithm (SVM), and a Multi-Layer Algorithm (ANN).

TABLE I: METRICS RESULTS FOR THE PROPOSED MODEL

	Random Forest	Support Vector Machines	Artificial Neural Network
Precision	0.73	0.83	**0.91**

As shown in Table I, the ANN is the best solution when you need to estimate an output based on two expected classes. Figure 6 shows the confusion matrix for this algorithm, it is another way to see the performance of our solution.

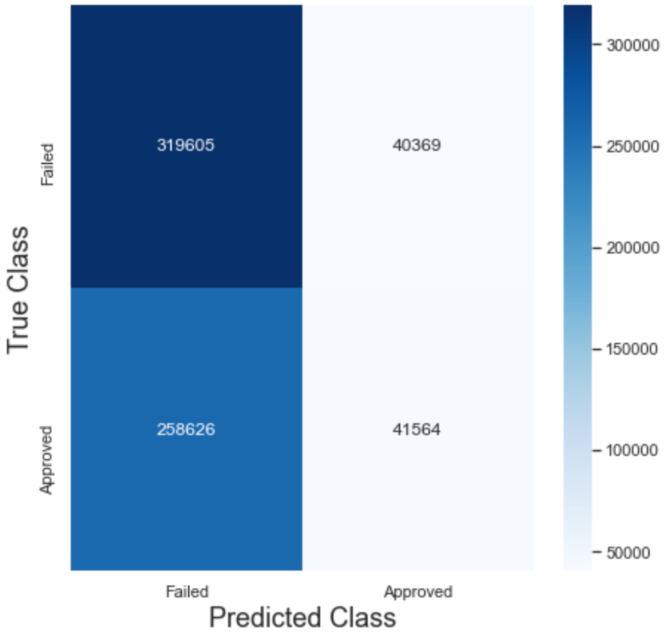


Figure 6: Confusion Matrix of the proposed ANN

V. CONCLUSIONS AND FUTURE WORKS

In this work we propose a framework for a course approval based on Portal Em Numeros data. Our results, an accuracy close to 91% is obtained when the ANN algorithm is applied for this task.

For future works, we expect to create an automatic data pipeline, using a tool called Airflow to automate our data extraction and allow us to improve the performance of our estimation and support the course coordinator to predict how successful the students will be.

ACKNOWLEDGMENT

This work is supported in part by CNPq – Brazilian National Research Council (No. PQ-2 312180/2019-5 on Cybersecurity No. 465741/2014-2), in part by the National Auditing Department of the Brazilian Health System SUS (No. DENASUS 23106.118410/2020-85), in part by the Brazilian Ministry of the Economy (Nos. DIPLA 005/2016 and ENAP 083/2016), in part by the Administrative Council for Economic Defense (No. CADE 08700.000047/2019-14), in part by the General Attorney of the Union (No. AGU 697.935/2019), and in part by the General Attorney's Office for the National Treasure (No. PGFN 23106.148934/2019-67) an the Federal District Research Support Foundation – FAPDF.

REFERENCES

[1] “Ipea - atlas,” in https://www.ipea.gov.br/atlasestado/consulta/75, 2019, pp. 1–4.

[2] “Portal em numeros,” in https://emnumeros.escolavirtual.gov.br/indicadores/, 2019, pp. 1–4.

[3] “Escola nacional de administração pública - ENAP,” in www.escolavirtual.gov.br, 2019, pp. 1–4.

[4] “Lei nº 12.527, de 18 de novembro de 2011 - lei de acesso a informação LAI.” in http://www.planalto.gov.br/ccivil03/ato2011−2014/2011/lei/l12527.htm, 2011, pp.1 − −4.

[5] V. C. G. Coelho, J. da Costa, D. A. da Silva, R. d. S. Júnior, F. L. de Mendonça, and D. G. Silva, “Mineração de dados educacionais no ensino a distância governamental,” Conferências Ibero-Americanas WWW/Internet e Computação Aplicada. Brasília, Brasil, pp. 77–84, 2016.

[6] E. Ferrarezi and J. A. Tomacheski, “Mapeamento da oferta de capacitação nas escolas de governo no brasil: gestão da informação para fortalecimento da gestão pública,” Revista do Serviço Público, vol. 61, no. 3, pp. 287–303, 2010.

[7] A. M. de Andrade, “Escolas de governo e seu papel no aperfeiçoamento do desempenho dos servidores públicos,” RECIMA21-Revista Científica Multidisciplinar-ISSN 2675-6218, vol. 2, no. 5, pp. e25 350–e25 350, 2021.

[8] T. P. C. Nascimento and A. Esper, “Evasão em cursos de educação continuada a distância: um estudo na escola nacional de administração pública,” Revista do Serviço Público, vol. 60, no. 2, pp. 159–173, 2009.

[9] “Ibm - spss,” in https://www.ibm.com/br- “Escola nacional de administração pública - ENAP,” in www.escolavirtual.gov.br, 2019, pp. 1–4.

[10] I. Corrêa, M. Camões, J. Meyer-Sahling, K. Mikkelsen, and C. Schuster, “Distorções de incentivo ao desempenho e redução de motivação no serviço público federal no brasil,” Revista do Serviço Público, vol. 71, no. 3, pp. 476–503, 2020.

[11] L. R. de Almeida, J. P. C. da Costa, R. T. de Sousa, E. P. de Freitas, E. D. Canedo, J. Prettz, E. Zacarias, and G. Del Galdo, “Motivating attendee’s participation in distance learning via an automatic messaging plugin for the moodle platform,” in 2016 IEEE Frontiers in Education Conference (FIE). IEEE, 2016, pp. 1–5.

[12] D. Ghosh and A. Vogt, “Outliers: An evaluation of methodologies,” in Joint statistical meetings, vol. 2012, 2012.

[13] M. Belgiu and L. Dragu¸t, “Random forest in remote sensing: A review of applications and future directions,” ISPRS journal of photogrammetry and remote sensing, vol. 114, pp. 24–31, 2016.

[14] de Oliveira Júnior, G.A.; de Oliveira Albuquerque, R.; Borges de Andrade, C.A.; de Sousa, R.T., Jr.; Sandoval Orozco, A.L.; García Villalba, L.J. Anonymous Real-Time Analytics Monitoring Solution for Decision Making Supported by Sentiment Analysis. *Sensors* 2020, *20*, 4557. https://doi.org/10.3390/s20164557

[15] E. S. Gualberto, R. T. De Sousa, T. P. De B. Vieira, J. P. C. L. Da Costa and C. G. Duque, "From Feature Engineering and Topics Models to Enhanced Prediction Rates in Phishing Detection," in *IEEE Access*, vol. 8, pp. 76368-76385, 2020, doi: 10.1109/ACCESS.2020.2989126.