Proposed Technological Solution to Predict the Need for Health Professionals in Health Centers Using Random Forest

Fiorella Patricia Mirano Surquislla[1], Gianfranco Henry Ore Paredes[2], Aguilar-Alonso Igor[3]*
Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional Mayor de San Marcos, Lima, Perú [1, 2]
Escuela Profesional de Ingeniería de Sistemas, Universidad Nacional Tecnológica de Lima Sur, Lima, Peru [3]*

Abstract—**The objective of this research is to develop a technological solution based on the Random Forest algorithm to predict healthcare workforce requirements in public healthcare centers in Peru, addressing staff shortages and unequal workforce distribution. A national dataset from the Peruvian Ministry of Health (MINSA) covering the period 2017–2024, segmented by levels of care (I, II, and III), was used to capture the operational differences within the healthcare system. The model, validated using an 80/20 split, achieved outstanding performance, with coefficients of determination (R^2) exceeding 0.99 and minimal percentage errors (MAPE) across all levels of care. The main contribution of this work lies in converting estimated healthcare attendances into an operational metric of "required healthcare professionals", integrated into a web-based architecture built on React, Flask, and PostgreSQL. The findings identify medical specialty and year as the most influential predictive variables. It is concluded that the proposed tool is robust for optimizing strategic healthcare workforce planning, enabling a more equitable and data-driven allocation of medical specialists.**

Keywords—Technological solution; Random Forest; healthcare sector; healthcare professional prediction; human resource management

I. INTRODUCTION

The shortage of healthcare professionals has been a persistent societal problem even before the onset of the COVID-19 pandemic. According to the World Health Organization (WHO), prior to 2020, a global shortage of approximately 18 million healthcare workers was estimated, with the deficit being more pronounced in countries with limited budgets and in those experiencing poor distribution of health resources, both human and material [1]. In Latin America, unequal workforce distribution, chronic occupational burnout, and inadequate planning for the recruitment and retention of medical teams have long-standing consequences for healthcare systems [2].

In Peru, the Office of the Comptroller General of the Republic has reported that nearly half of public healthcare facilities do not meet the minimum required number of professionals to provide adequate care, revealing a serious shortage of physicians, nurses, and technical staff [3]. Additionally, a technical report from the Institute for Health Technology Assessment and Research (IETSI) of EsSalud (2024) highlights critical gaps in both the quantity and distribution of healthcare personnel across regions and levels of care [4]. These gaps underscore the urgent need to implement predictive systems that enable more accurate and equitable healthcare workforce planning.

At the global level, the WHO (2021) reported that many countries are facing severe shortages of healthcare professionals and projected that millions of additional workers will be required by 2030, particularly in low-resource settings [5]. However, although some governments have expanded healthcare coverage through insurance schemes or public programs, this does not necessarily ensure equitable access to care. In Ghana, for example, individuals with better socioeconomic conditions have been shown to benefit more from such programs than more vulnerable populations [6]. The COVID-19 pandemic further exacerbated these challenges by exposing long-standing structural weaknesses in Latin American healthcare systems, such as poor workforce distribution, sustained occupational fatigue, and insufficient workforce planning—issues also documented in studies conducted in Brazil [7]. Consequently, current strategies must move beyond merely increasing budgets or coverage and instead focus on improving workforce organization, distribution, and proactive planning to ensure equitable access to healthcare services, particularly in high-demand and resource-constrained areas.

At the regional level, increasing attention has been paid in recent years to evaluating the efficiency of public healthcare systems, especially in Latin America. Studies based on Data Envelopment Analysis (DEA) have revealed substantial differences in system performance across countries, even when public health expenditure levels are comparable [8]. For this reason, assessing efficiency in Primary Healthcare (PHC) has become particularly relevant. Even in contexts with broad coverage and sufficient staffing—such as Cuba—significant regional disparities persist, often associated with population size and local socioeconomic conditions [9]. These findings demonstrate that increased financial investment alone does not guarantee improved outcomes; rather, the effective use of available resources plays a crucial role. Identifying underperforming healthcare centers and understanding the factors influencing their performance can support better decision-making, enhance management practices, and ensure that resources translate into improved healthcare delivery and population health—an essential requirement for building sustainable health systems in the region [10].

Similarly, in countries such as Indonesia, where healthcare workforce shortages—particularly of physicians—are critical, unequal distribution has negatively affected primary healthcare performance. In the face of such disparities and limited

*Corresponding author.

resources, it is essential to evaluate how efficiently healthcare facilities utilize available personnel to deliver key services. This assessment enables optimal resource use and supports the estimation of potential gains in service coverage if the existing workforce is deployed more effectively [11].

Given the difficulty of organizing healthcare staff without precise data on future demand, this study poses a central research question: To what extent does the development of a technological solution based on the Random Forest algorithm allow for the accurate prediction of the need for healthcare professionals in Peru's public centers, considering the variability across levels of care? The hypothesis suggests that the integration of geographic, temporal, and specialty variables, segmented by levels (I, II, and III), offers superior reliability compared to conventional statistical methods. The objective is to develop a predictive model capable of accurately estimating the required specialists, recognizing that each level possesses distinct demand dynamics and resolution capacities.

The differential value of this work is built upon three technical pillars. First, it utilizes a massive database from the Ministry of Health (MINSA) with seven years of national records (2017–2024), capturing both regular seasonality and the impact of the health crisis. Second, the analysis is rigorously segmented into the three levels of primary care, acknowledging that resolution capacity varies drastically according to the complexity of the facility. Finally, this proposal translates appointment volume into a "required professionals" metric through an operational conversion methodology, supported by a complete software architecture based on React, Flask, and PostgreSQL

The structure of the manuscript is organized as follows: Section II presents the related work, including a brief review of the relevant literature. Section III details the methodological framework, describing data preprocessing, feature engineering, and the training of the Random Forest algorithm, as well as the technical definition of the evaluation metrics. Section IV presents the architecture of the proposed technological solution. Section V reports the results based on the accuracy metrics obtained for each level of care, along with the analysis of feature importance and regional demand. Section VI discusses the results, which are contrasted with the scientific evidence identified in the literature review. Finally, Section VII summarizes the conclusions and outlines directions for future work.

II. Related Work

Recent research has applied various machine learning techniques to predict the demand for healthcare services, utilizing models that incorporate demographic, economic, and health-related factors.

A study based on Andersen's Behavioral Model of Health Services Use analyzed 2022 data from Turkey [12], identifying variables such as gender, age, educational level, treatment costs, and the presence of chronic diseases as significant predictors of healthcare utilization. Models such as Random Forest, XGBoost, and Logistic Regression demonstrated high accuracy and generalization capabilities in predicting total service demand, providing valuable insights for the efficient allocation of resources in public health centers. These findings highlight the potential of machine learning algorithms to enhance decision-making in public health policy through data-driven analysis.

Furthermore, another study [13] proposed a web-based system using machine learning to mitigate the shortage of healthcare professionals in Peru through Electronic Health Records (EHR). This system implemented the XGBoost algorithm to predict demand across different medical specialties, enabling optimized staff planning. The model achieved a 93% efficiency rating in the AUC test, identifying medical services with the highest projected demand. The automation of data analysis significantly reduced the administrative burden on staff, freeing up time for patient care activities. However, reliance on historical data from a single health center and the lack of standardization in medical record formats limited the system's immediate scalability at a national level.

Yadav [14] investigated the application of machine learning models to optimize professional allocation in healthcare systems with fluctuating demand. The study evaluated four algorithms: Logistic Regression, Decision Tree, Random Forest, and Support Vector Machines (SVM), using historical staffing schedules and patient records. The Random Forest model achieved the highest performance with 98% accuracy, followed by Decision Tree (97%), Logistic Regression (96%), and SVM (95%). The study demonstrated that these models allow for the anticipation of staffing needs and better shift organization by considering both predicted demand and personnel competencies. Nevertheless, a noted limitation is the current lack of tools that combine real-time demand forecasting with automated scheduling. Collectively, these results suggest that machine learning can improve institutional internal operations and reduce excessive workloads.

In the context of human resource planning, a study [15] demonstrated the applicability of machine learning models to predict the retention of healthcare professionals in rural communities in South Africa. Utilizing demographic data from health workers, they trained three classification models (Multinomial Logistic Regression, Decision Tree, and Naïve Bayes), which reached a predictive accuracy of 47.34% and an average AUC of 0.66, significantly outperforming random chance. This study emphasizes the potential of these techniques to optimize recruitment and retention processes in hard-to-reach areas, while also pointing out significant limitations regarding data quality and completeness in resource-constrained environments.

Koç and Eren [16] conducted research to project the availability of physicians in Turkey through the year 2030. To achieve this, they compared various machine learning and deep learning algorithms, including XGBoost, LSTM, and Random Forest, using Root Mean Square Error (RMSE) to measure their accuracy. The study combined historical information from the OECD with artificially generated data to strengthen the analysis of the diverse demographic and economic factors influencing the sector. Following the testing phase, the XGBoost model proved to be the most accurate in its predictions. The final results suggest that a professional deficit will exist by 2030, with an estimated supply of 3.04 physicians per 1,000 inhabitants against a demand of 3.12, underscoring the need to revise current healthcare training policies.

Köksoy et al. [17] developed a study to predict the need for healthcare personnel across the 81 provinces of Turkey, aiming to reduce inequality in its distribution. They compared nine artificial intelligence models, highlighting the use of the Random Forest Regressor, which achieved a high level of precision with an R^2 of 0.99 and a Mean Absolute Error (MAE) of 39.54. These figures indicate that the model is capable of explaining almost the entire variance of the data and that its margin of error is minimal when estimating the number of professionals. By analyzing data from 2002 to 2023, the authors successfully projected 2024 demands with high accuracy, demonstrating that these algorithms are key tools for more efficient hospital planning.

III. Methodology

This research is applied in nature, as it seeks not only to describe the current state of the healthcare sector but also to propose a practical solution for implementation. Specifically, it involves the development of a technological solution using the Random Forest machine learning algorithm to predict the future demand for public healthcare professionals.

The research level is predictive, as it aims to forecast future healthcare professional demand based on historical data. Through the machine learning model, the study expects to anticipate personnel shortage scenarios commonly found in Peru, where demand in highly populated regions often exceeds the available healthcare workforce. This will enable data-driven strategic decision-making and, potentially, improve the distribution of physicians across the country.

The research is oriented toward healthcare professionals working within the public system, specifically across primary, secondary, and tertiary levels of care managed by the Ministry of Health (MINSA). These categories include specialized physicians, nurses, obstetricians, technicians, and other staff. This topic was selected because it addresses a critical contemporary issue: the deficiency and inequitable distribution of healthcare professionals in the public sector, which directly impacts the quality of service provided to the population.

A. Data Sources and Dataset Description

For the development of the predictive model, the dataset titled Healthcare Attendances Provided to Insured Patients – [SIS] was used. This dataset is provided by the Peruvian Ministry of Health (MINSA) and published on the National Open Data Platform, covering the period from 2017 to June 2024. Its selection is justified by its nationwide coverage and the depth of a seven-year historical record, which enables the model to accurately capture long-term trends, seasonal patterns, and anomalies derived from the healthcare crisis.

1) The dataset includes the following columns:

- YEAR, MONTH, REGION, PROVINCE, UBIGEO_DISTRICT, DISTRICT, EXECUTING_UNIT_CODE, EXECUTING_UNIT_DESCRIPTION, IPRESS_CODE, IPRESS, HEALTHCARE_LEVEL, INSURANCE_PLAN, SERVICE_CODE, SERVICE_DESCRIPTION, SEX, AGE_GROUP, ATTENDANCES.

B. Data Preprocessing and Feature Engineering

Once the required datasets were obtained to initiate the development of the artifact, the process began with data cleaning and transformation of the original data.

1) Data preprocessing: Before training the model, a data cleaning and preprocessing process was performed, which included removing missing values, correcting inconsistencies, and standardizing variable formats.

The first preprocessing step focused on data validation and cleaning. Initially, records with missing values in critical variables defining each observation (YEAR, MONTH, IPRESS_CODE, SPECIALTY, and geographic variables) were removed, ensuring the integrity of the analyzed records. Subsequently, to mitigate the influence of extreme values during training, a robust outlier filtering approach was applied: observations of the target variable, ATTENDANCES, were bounded within the interpercentile range P1–P99. After cleaning, the data were aggregated by summing total ATTENDANCES at the most detailed spatio-temporal unit (YEAR, MONTH, IPRESS_CODE, SPECIALTY, REGION, PROVINCE, UBIGEO_DISTRICT), consolidating the information for analysis.

Prior to training, nominal categorical variables (IPRESS_CODE, SPECIALTY, REGION, PROVINCE, UBIGEO_DISTRICT) were converted into numerical format using the Label Encoding method. Finally, the target variable ATTENDANCES, which exhibited a skewed distribution, was subjected to a logarithmic transformation, which is explained in a later section.

2) Feature engineering: New features were created through feature engineering techniques, which involve transforming and enriching the original data to capture patterns that are not directly observable.

3) Temporal features: This group incorporates time-related information to capture seasonal and cyclical patterns affecting demand. It includes: SEASON, IS_END_OF_YEAR, IS_START_OF_YEAR, IS_PEAK_MONTH, MONTH_SIN, MONTH_COS.

4) Geographical features: This set groups variables that describe the territorial context and the distribution of services within each geographical area. It consists of: IPRESS_DENSITY_REGION, IPRESS_DENSITY_PROVINCE, SPECIALTY_FREQUENCY.

5) Historical features: This block includes features derived from past behavior to capture trends, variability, and consistency in the data. It includes: HISTORICAL_MEAN, HISTORICAL_STD, HISTORICAL_MAX, HISTORICAL_MIN, HISTORICAL_COUNT, REGION_SPECIALTY_MEAN, VARIATION_VS_MEAN, CAPACITY_RATIO.

These features were created from the following original variables (IPRESS, specialty, region, province, month, year) to compute averages, maxima, minima, and count-based metrics.

C. Algorithm and Training

1) Algorithm: The selection of the Random Forest algorithm for this study is based on its well-recognized predictive capability in the healthcare domain. A recent study comparing several machine learning models for pressure ulcer prediction showed that Random Forest achieved the best performance, with an accuracy of 99.88% and an AUC value of 0.9999, significantly outperforming other algorithms such as SVM, Decision Tree, and ANN [16]. This superiority is attributed to its ability to handle multiple predictive variables, reduce overfitting through the aggregation of multiple decision trees, and capture complex non-linear relationships within the data. For these reasons, Random Forest is particularly suitable for risk prediction in clinical contexts involving multifactorial data.

The developed model consists of two main steps:

a) Prediction of healthcare attendances.
b) Calculation of required healthcare professionals.

First, healthcare attendances are predicted. For this purpose, the model was configured with parameters including 300 trees (n_estimators = 300) and a maximum depth of 25 levels (max_depth = 25), allowing the model to capture complex patterns while controlling overall complexity. Bootstrap sampling with 80% of the data (max_samples = 0.8) was applied, along with random selection of √n features at each split to increase ensemble diversity.

To handle the uneven distribution typically observed in healthcare attendance data—where low values predominate but some very high values exist—a logarithmic transformation was applied prior to training. This technique compresses the data scale, enabling the model to learn patterns from both small and large healthcare facilities. The predictions were subsequently transformed back to their original scale, resulting in more balanced and accurate estimates across the full range of healthcare establishments.

The second step consists of calculating the number of healthcare professionals required to meet the projected demand. For this purpose, the following variables were defined:

- PHN: Required healthcare professionals.
- AE: Estimated monthly attendances.
- R: Professional productivity (attendances per hour).
- JH: Effective working hours (total contracted hours per year).

The relationship among these variables is expressed in Eq. (1):

$$\mathrm{PHN} = \max(1, round\,(\frac{\mathrm{AE}}{\mathrm{JH\ x\ R}})) \quad (1)$$

To obtain the average number of monthly attendances across the different levels of healthcare centers, the Programming Criteria Standards table [19] was used. From this source, the following parameters were obtained:

- Level I: 500 attendances/month per professional
- Level II: 277.6 attendances/month per professional
- Level III: 270 attendances/month per professional

Likewise, the Maximum (max) function ensures a minimum of one professional per healthcare facility, guaranteeing basic coverage even in locations with low demand.

2) Data Split – Training: The complete set of clean and preprocessed data (X for the features and ATTENDANCES as the target variable) was divided into two main subsets through a random and controlled process:

- Training Set (80%): The majority of the data, 80%, was used to train the model. This allowed the algorithm to learn the underlying patterns and relationships between the features (geographical, temporal, and historical) and the target variable.
- Test Set (20%): The remaining 20% of the records were reserved. This subset was not exposed to the model during the learning phase. Its sole purpose was to measure the model's performance and accuracy on completely unseen data, ensuring an unbiased evaluation of its generalization capability.

To ensure consistency in this partitioning process and allow reproducibility of the results in future analyses, a fixed random seed (random_state = 42) was applied during data splitting.

D. Evaluation and Metrics

The model was independently evaluated for each level of the Peruvian healthcare system (I, II, and III), taking into account the operational characteristics specific to each level:

- Level I: High-frequency, low-complexity healthcare attendances
- Level II: Medium-complexity and specialized healthcare attendances
- Level III: High-complexity, low-frequency healthcare attendances

To assess performance at each level, four key metrics were used:

The MAE (Mean Absolute Error), defined in Eq. (2), calculates the average absolute error between the actual values and the predictions, providing a direct measure of the error expressed in the same unit as the attendances (number of attendances per month):

$$\mathrm{MAE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}|\mathrm{y_i} - \hat{y_\imath}| \quad (2)$$

where, y_i represents the actual value, $\hat{y}_i$ the model prediction, and n the number of observations in the test set. This metric quantifies the average error in the number of attendances, which is directly related to healthcare workforce planning.

The RMSE (Root Mean Squared Error), defined in Eq. (3), penalizes large errors more heavily than small ones, making it particularly useful for identifying predictions with significant deviations.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

The R^2 (Coefficient of Determination), defined in Eq. (4), quantifies the proportion of variance in the dependent variable that is explained by the model, indicating how well the predictions fit the actual data.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

where, $\bar{y}$ represents the mean of the actual values. R^2 quantifies the percentage of variance in the dependent variable that is explained by the model. This metric evaluates the model's ability to explain variability in healthcare service demand.

The MAPE (Mean Absolute Percentage Error), defined in Eq. (5), expresses the error as a percentage, facilitating the interpretation of model performance in relative terms. This is particularly important when comparing different levels of healthcare delivery.

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{\max\,(y_i, 1)}\right| \tag{5}$$

This metric provides the relative error, allowing comparisons across different levels of healthcare delivery.

Additionally, several analyses were conducted to gain deeper insight into the model's behavior. The most influential features in the predictions were evaluated in order to identify which variables are determinant in demand estimation. The medical specialties with the highest workforce requirements were also analyzed, making it possible to identify those with the largest staffing gaps. Finally, results were compared across regions to determine where the greatest need for healthcare personnel exists, revealing significant territorial differences. For all these evaluations, explanatory graphs were used, which were generated from the model's results.

IV. Proposed Solution

To address the problem of healthcare workforce requirements in public healthcare centers in Peru, the design and implementation of a technological solution with an interactive and intuitive interface is proposed. This solution allows the user to select a healthcare center and obtain information on the number of physicians required, according to the specialties available at that facility in Peru.

Fig. 1 illustrates how the web application is integrated with the predictive model and the database. On the frontend, the user accesses a dashboard where a healthcare center, medical specialty, month, and year can be selected. This information is sent to the backend through an API built with Flask, which receives the request and queries the data stored in PostgreSQL. Based on this information, the system executes the trained model and generates the corresponding prediction. Finally, the result is returned to the frontend, where it is clearly displayed so that the user can easily interpret it.

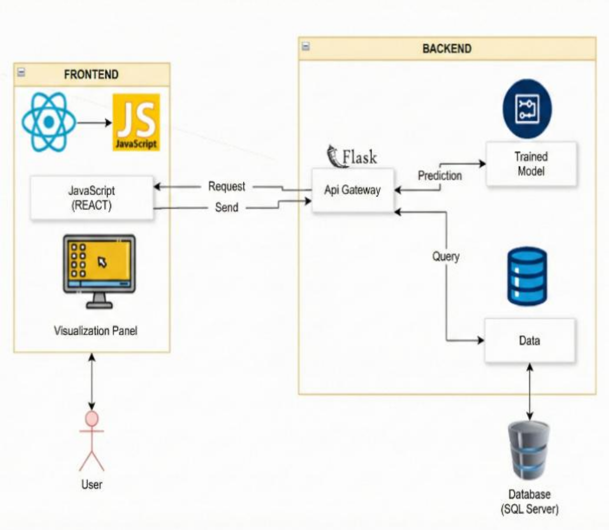
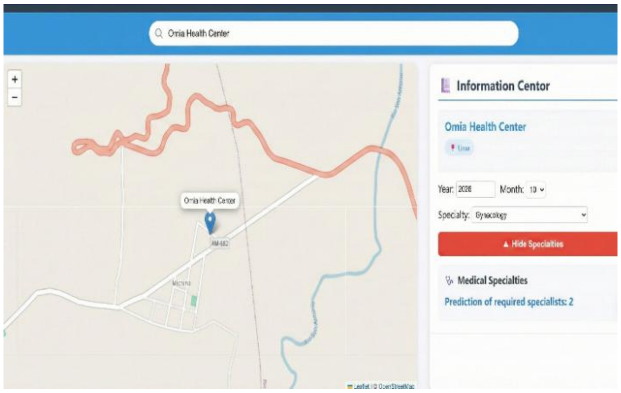


Fig. 1. Proposed technological solution.

The interface component of the project was developed using React.js, a widely used and well-established JavaScript framework for building web applications. React is based on principles of speed, security, and ease of integration with other libraries, making it a suitable and non-complex choice for interface development.

The frontend of the application was designed to provide a user-centered experience. One of its main functionalities is the generation of an interactive view in which users can visualize the number of healthcare centers and their corresponding medical workforce requirements. This information is represented through an interactive mini-map, enabling a more intuitive and visual identification of areas with greater demand for healthcare professionals. The goal is to ensure that any user of the application can access the information in a simple, fast, and understandable manner. Additionally, the use of a map provides an effective way to understand the situation of healthcare centers, as shown in Fig. 2.



Omia Health Center

Information Centor

Omia Health Center

Year: 2026 Month: 10

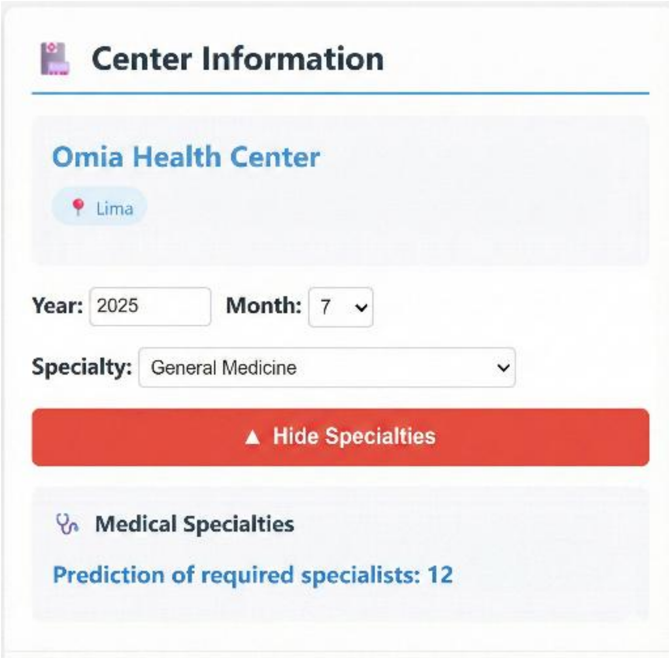
Specialty: Gynecology

Medical Specialties

Prediction of required specialists: 2

Fig. 2. Main screen.

The main sidebar includes filters such as year, month, and medical specialty. Based on these selections, the predicted number of required healthcare professionals per specialty is displayed for each healthcare center selected through the interactive map or the bar-based search interface, as shown in Fig. 3.



Center Information

Omia Health Center

Lima

Year: 2025 Month: 7

Specialty: General Medicine

Medical Specialties

Prediction of required specialists: 12

Fig. 3. Main sidebar.

The interface presents the model predictions in a more structured manner, filtering them by time period and medical specialty. A REST API is used as the communication channel between the frontend and the backend. The frontend issues requests to retrieve the results generated by the predictive models, while the backend processes these requests, queries the prediction model, and returns the response in a clear and standardized format—such as JSON—so that the results can be easily rendered within the application.

V. Results

The implemented model demonstrated a high level of accuracy in predicting the demand for medical professionals across the three levels of primary healthcare in Peru. The results show that the algorithm maintained robust performance in all evaluated scenarios, with minimal prediction errors (MAPE ranging between 1.28% and 6.17%) and a consistently high explanatory power ($R^2 > 0.99$). Feature importance analysis identified medical specialty as the most influential factor, followed by the year and geographic location variables.

The analysis of the obtained results is presented across two fundamental dimensions. First, Section A details the performance metrics and statistical validation that support the reliability of the model. Subsequently, Section B analyzes the key predictive factors and presents specific case studies that demonstrate the operational applicability of the system in real healthcare facilities.

A. Performance Metrics and Validation

Model validation was performed using the hold-out data splitting technique, where 80% of the records were used for training, and the remaining 20% were reserved as the test set. As detailed in Table I, Level I exhibits minimal errors (MAE = 0.27), while at Level II the model maintains excellent explanatory capability ($R^2 = 0.9968$).

For Level III, despite the higher volume of healthcare services, the model's performance remains strong, achieving an R^2 value of 0.9938.

TABLE I. Metrics Performance

Level	*MAE*	*RMSE*	R^2	*MAPE*
I	0.265715	1.202290	0.995892	1.284139
II	8.838402	32.882899	0.996812	3.455243
III	48.685679	183.381062	0.993822	6.168446

Additionally, to verify the absence of overfitting and the reliability of the estimates, scatter plots were generated to compare the observed values against those predicted by the algorithm.

As shown in Fig. 4, Level I exhibits a high density of points aligned along the ideal diagonal ($R^2 = 0.9959$). However, the dispersion increases as the values grow, indicating high confidence in predictions at lower demand levels, which gradually decreases for higher values. Nevertheless, it is important to note that most of the high-value observations at this level correspond to atypical cases.

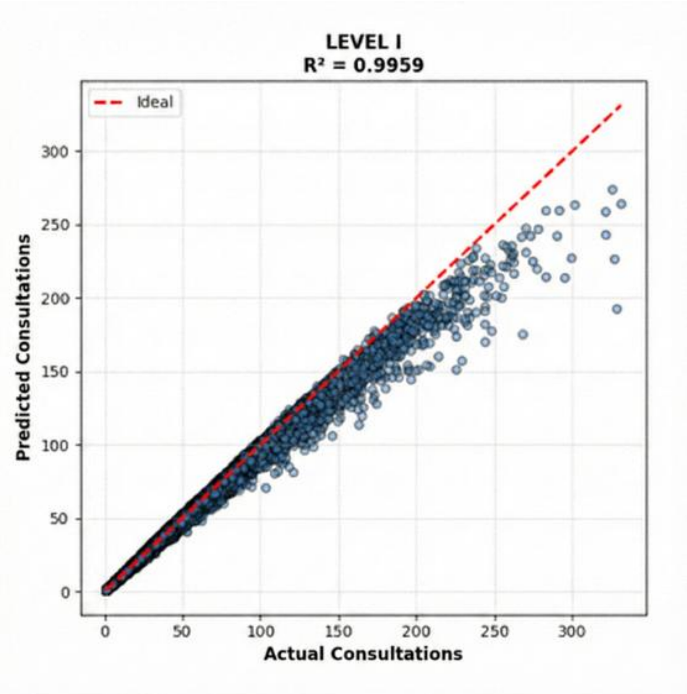


Fig. 4. Scatter Plot – Level I.

As shown in Fig. 5, Level II exhibits a consistently high density of points along the ideal diagonal, with an R^2 value of 0.9968. Only very slight variations are observed when compared to the Level I scatter plot. This behavior indicates a high level of model stability and, consequently, a high degree of confidence in the predictions.

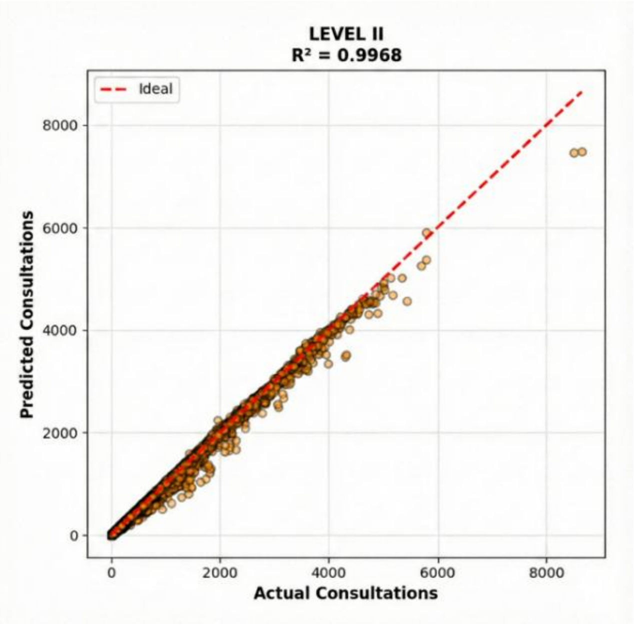


Fig. 5. Scatter Plot – Level II.

In Fig. 6, Level III shows a consistently high density of points along the ideal diagonal, with an R^2 of 0.9968, exhibiting only very slight variations compared to the Level I plot. This behavior indicates high model stability and, consequently, strong confidence in the predictions.

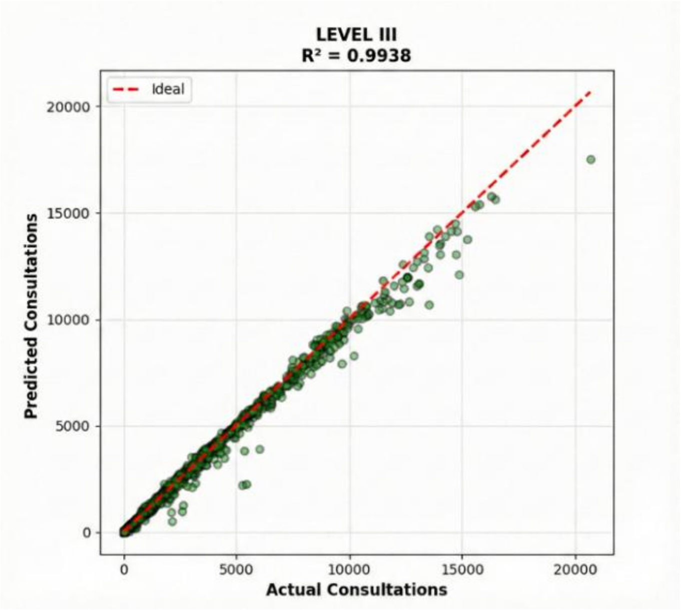


Fig. 6. Scatter Plot – Level III.

As observed in the figures for Level I (Fig. 4), Level II (Fig. 5), and Level III (Fig. 6), the results exhibit an almost linear fit. However, in Levels I and III, observations with higher volumes of healthcare attendances tend to fall slightly below the reference line. This minor underestimation at the extremes of demand is minimal and does not compromise the overall reliability of the tool, which maintains accuracy levels consistently above 99% across all evaluated scenarios.

B. *Importance of Characteristics in Predicting Healthcare Personnel*

Once the statistical robustness of the model was confirmed, the factors determining healthcare workforce demand in the Peruvian public health system were analyzed. The feature importance analysis reveals a clear hierarchy in the influence of predictive variables across the three evaluated levels of care.

Fig. 7 illustrates the top features that most influenced the prediction, highlighting Specialty (0.2%) and Year (0.15%) as the most influential features, while Province (0.04%) and Month (0.03%) showed the lowest influence.

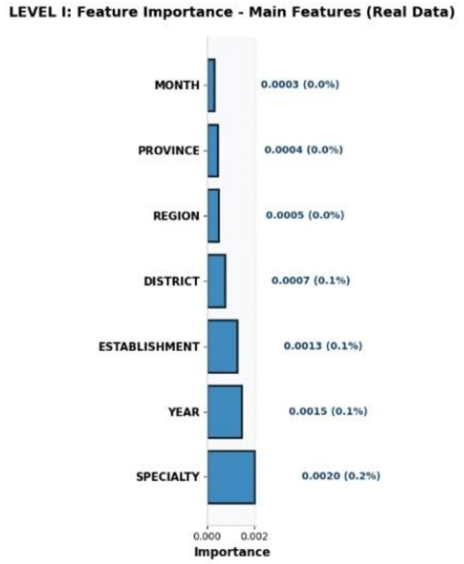


Fig. 7. Most influential main features – Level I.

Fig. 8 shows the top features that most influenced the prediction, highlighting Specialty (0.88%) and Year (0.42%) as the most influential features, while Region (0.13%) and Month (0.07%) were the least influential.

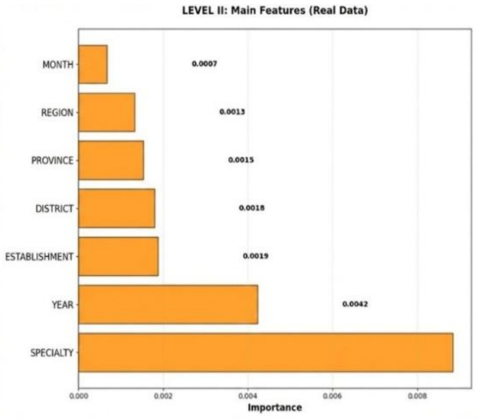


Fig. 8. Most influential main features – Level II.

Fig. 9 shows the top features that most influenced the prediction, highlighting Specialty (1.37%) and Year (0.62%) as the most influential features, while Province (0.14%) and Month (0.01%) were the least influential.

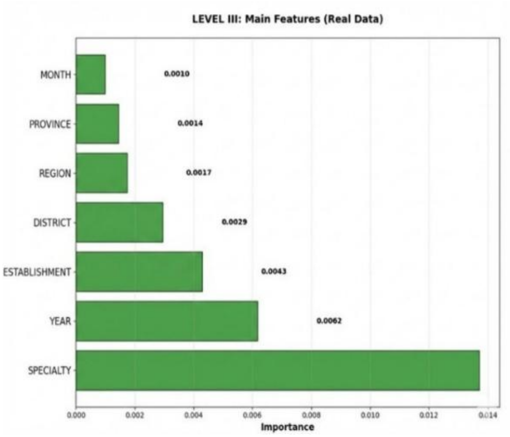


Fig. 9. Most influential main features – Level III.

In all cases, the SPECIALTY variable is consolidated as the most critical and determining factor in the model's behavior. In a second order of relevance, the YEAR variable appears, indicating a significant temporal progression in demand, although its relative weight is considerably lower than that of the medical specialty. On the other hand, geographic variables such as FACILITY, DISTRICT, and PROVINCE show a moderate influence, while REGION and MONTH represent the least influential predictive factors for the algorithm.

Finally, the analysis by medical specialty makes it possible to prioritize areas with the greatest shortage of human resources. For this purpose, prediction tests were conducted on randomly selected health centers for the period of August 2023. As shown in Fig. 10, in the Level I health center with code 381, the highest number of attendances corresponded to the General Medicine specialty, which is why a specialist in this area would be prioritized. In addition, it is observed that the difference between actual and estimated attendances presents only minimal decimal variations.

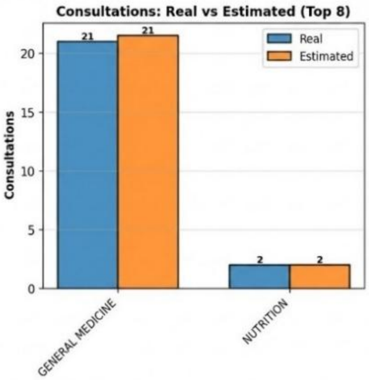


Fig. 10. Top specialties by attendance – Level I.

In Fig. 11, it is observed that in the Level II health center with code 5366, the highest number of attendances corresponded to the Nursing specialty. For this reason, a specialist in this area would be prioritized. Additionally, the variation between the number of actual and estimated attendances for this specialty is 6, which represents a minimal difference that does not affect the reliability of the prediction results.

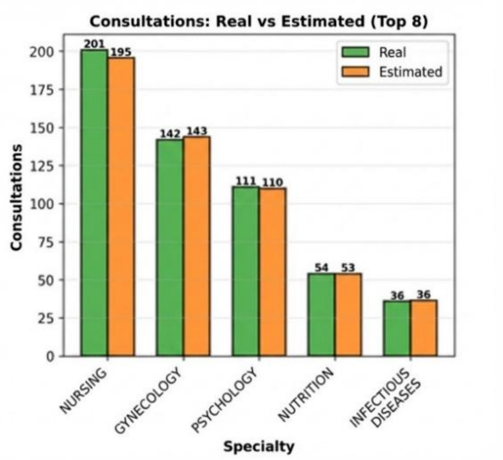


Fig. 11. Top specialties by attendance – Level II.

In Fig. 12, it is observed that in the Level III health center with code 2289, the highest number of attendances corresponded to the Pediatrics specialty. For this reason, a specialist in this area would be prioritized. Additionally, the variation between the actual and estimated attendances is 8, which does not directly affect the prediction results.

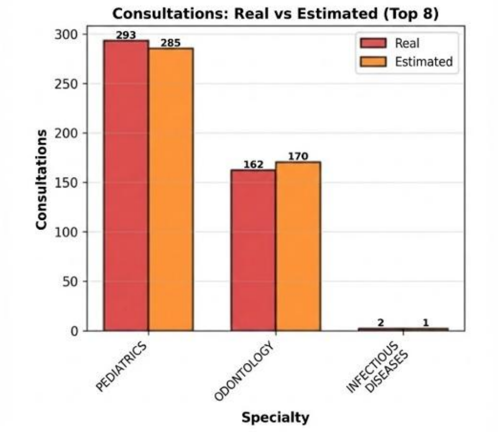


Fig. 12. Top specialties by attendance – Level III.

Overall, these findings allow us to conclude that the technological solution accurately identifies not only how many health professionals are required, but also in which specific areas urgent intervention is needed. This capability to segment results by levels of care and medical specialties constitutes a fundamental input for human resource management, enabling the reduction of care gaps and the optimization of the problem-solving capacity of public health centers in Peru.

VI. Discussion

The discussion of the findings is structured around two main axes: predictive robustness in relation to international standards and the relevance of the identified variables for human resource management.

A. Comparative Analysis of Predictive Performance

The performance of the Random Forest algorithm in this study, with an $R^2 > 0.99$ across all three levels of care, aligns with the highest global standards of predictive accuracy. When compared with the literature, these results match the coefficient of determination reported by Köksoy et al. [18] in Turkey, while demonstrating a significant competitive advantage in terms of local error margins. Whereas Köksoy et al. reported an MAE of 39.54, Level I in this study achieved an MAE of 0.27, evidencing superior precision for primary care demand. Likewise, the accuracy obtained exceeds the 0.98 reported by Yadav [14], confirming that the model's fit is optimal for the complexity of the Peruvian healthcare system.

In terms of robustness, the superiority of this model over the 47.34% accuracy and 0.66 AUC reported by Moyo et al. [15] suggests that the inclusion of geographic and temporal variables from MINSA enables substantially higher predictive fidelity. Even at Level III, the model maintained a MAPE of 6.17%, a remarkably low figure compared to the high-specialization projections of Koç and Eren [16], which relied on synthetic data. The use of seven years of real historical records in this study ensures that the reported metrics reflect a proven capacity for generalization under fluctuating demand scenarios.

B. Factor Importance and Contribution to Healthcare Management

The feature importance analysis identifies SPECIALTY and YEAR as the most influential factors in the model, enabling strategic planning based on medical complexity and temporal progression. Unlike the behavioral approaches of Orhan and Kurutkan [12], which predict service utilization intentions based on personal factors, this proposal introduces a direct operational management metric. Through the technical conversion formula, the approach goes beyond appointment volume forecasting to generate an actionable output: the number of "required health professionals".

This operational capability addresses the lack of scheduling and planning tools identified by Yadav [14] and surpasses the scalability of local studies such as Perez-Siguas et al. [13], whose 93% AUC efficiency was confined to specific environments. Validation through real-world cases (health centers 381, 5366, and 2289) demonstrates that the differences between actual and estimated demand are minimal, enabling equitable resource allocation. Finally, integration into a React- and Flask-based architecture ensures that the achieved accuracy is not merely a theoretical result, but rather a scalable and sustainable administrative decision-support tool for the Peruvian public healthcare system.

VII. Conclusion

The present study demonstrates that the development of a technological solution based on the Random Forest algorithm enables highly accurate prediction of healthcare workforce requirements in public healthcare centers in Peru, thereby validating the hypothesis that the integration of geographical, temporal, and specialty-related variables—segmented by levels of care—significantly improves predictive performance compared to traditional statistical approaches. The model achieved a coefficient of determination $R^2 > 0.99$ across all levels of care, with low percentage errors (MAPE ranging from 1.28% for Level I to 6.17% for Level III), evidencing robust performance even in high-complexity scenarios. In addition, medical specialty and temporal evolution were identified as the most influential variables, enabling more targeted and informed workforce allocation decisions.

The main competitive advantage of this research lies in its ability to transform projected healthcare service demand predictions into operational management metrics for the required number of healthcare professionals, thereby facilitating administrative decision-making and surpassing prior approaches that focused solely on forecasting appointment volumes or patient retention behaviors with lower accuracy. The implementation of the model within a functional web-based architecture further confirms its technical feasibility as a decision-support tool for healthcare management.

Nevertheless, critical limitations related to the data source are acknowledged. Reliance on open data from the Peruvian Ministry of Health (MINSA) introduces vulnerability to underreporting in rural areas with limited connectivity, which may bias projections for Level I healthcare centers with constrained infrastructure. Moreover, the dataset does not map all service areas within healthcare centers, but only attendances provided and covered by the Comprehensive Health Insurance (SIS). Additionally, generalizing the model to contexts outside the Peruvian public healthcare system would require retraining with exogenous variables that account for different cost structures and patient flow dynamics.

As future work, the integration of real-time epidemiological and climatic variables is recommended to strengthen the model's resilience to disease outbreaks. Finally, validation of this solution in real-world hospital management applications is proposed to assess its direct impact on reducing patient waiting times, thereby contributing to the consolidation of a data-driven and technologically efficient healthcare system.

References

[1] "Datos acerca de los trabajadores sanitarios y asistenciales," OMS | Organización Mundial de la Salud. https://www.who.int/es/campaigns/annual-theme/year-of-health-and-care-workers-2021/facts.

[2] Portela, G. Z., Fehn, A. C., Ungerer, R. L. S., & Poz, M. R. D. "Human resources for health: global crisis and international cooperation. Recursos humanos em saúde: crise global e cooperação internacional", Ciencia & saude coletiva, vol22, no. 7, pp. 2237–2246, Jul 2017, doi: 10.1590/1413-81232017227.02702017.

[3] Contraloría General de la República. La mitad de las postas médicas no cuentan con personal de salud mínimo para atender pacientes. 2016.

[4] P. C. Cecilia, C. P. E. Juan, P. I. Daniel, S. B. Percy, and A. B. Fabio, "Desarrollo y Validación de Nueva Metodología para Determinar la Brecha de Profesionales Médicos Según el Perfil Epidemiológico de la Demanda de Consulta Externa en las IPRESS con Población Adscrita de EsSalud: Propuesta y Caso Aplicado a la Red Prestacional Rebagliati. Reporte de Resultados de Investigación 07-2024," 2024. https://hdl.handle.net/20.500.12959/5199.

[5] "Nuevo informe de la OPS revela que 14 países de las Américas enfrentan escasez de trabajadores de salud," OPS/OMS | Organización Panamericana De La Salud. https://www.paho.org/es/noticias/30-4-2025-nuevo-informe-ops-revela-que-14-paises-americas-enfrentan-escasez-trabajadores.

[6] M. Boniol, T. Kunjumen, T. S. Nair, A. Siyam, J. Campbell, and K. Diallo, "The global health workforce stock and distribution in 2020 and 2030: a threat to equity and 'universal' health coverage?," BMJ Global Health, vol. 7, no. 6, p. e009316, Jun. 2022, doi: 10.1136/bmjgh-2022-009316.

[7] G. Z. Portela, A. C. Fehn, R. L. S. Ungerer, and M. R. D. Poz, “Recursos humanos em saúde: crise global e cooperação internacional,” Ciência & Saúde Coletiva, vol. 22, no. 7, pp. 2237–2246, Jul. 2017, doi: 10.1590/1413-81232017227.02702017.

[8] M. Yaghoubi, M. Salimi, and M. Meskarpour-Amiri, “Systematic review of productivity loss among healthcare workers due to Covid-19,” The International Journal of Health Planning and Management, vol. 37, no. 1, pp. 94–111, Oct. 2021, doi: 10.1002/hpm.3351.

[9] A. E. O. García, “Desigualdad en la distribución de médicos en el Perú,” Revista Cubana de Salud Pública, vol. 47, no. 1, Dec. 2020, [Online]. Available: http://www.revsaludpublica.sld.cu/index.php/spu/article/download/1447/1610.

[10] E. Espinoza-Portilla, W. Gil-Quevedo, and E. Agurto-Távara, “Principales problemas en la gestión de establecimientos de salud en el Perú,” Revista Cubana de Salud Pública, vol. 46, no. 4, Dec. 2020, [Online]. Available: http://scielo.sld.cu/pdf/rcsp/v46n4/1561-3127-rcsp-46-04-e2146.pdf.

[11] A. Niedar, F. Hafidz, and K. Hort, “Optimization of healthcare workers availability: Increasing primary health care efficiency in Indonesia,” Jurnal Ekonomi Kesehatan Indonesia, vol. 7, no. 1, p. 1, Jul. 2022. doi: 10.7454/eki.v7i1.5397.

[12] Orhan, F., Kurutkan, M.N. Predicting total healthcare demand using machine learning: separate and combined analysis of predisposing, enabling, and need factors. BMC Health Serv Res 25, 366 (2025). 10.1186/s12913-025-12502-5.

[13] R. Perez-Siguas, H. Matta-Solis, E. Matta-Solis, H. Matta-Perez, L. Perez-Siguas, and Z. O. Pumacayo-Sanchez, “Implementation of machine learning to mitigate the deficit of health personnel and optimize healthcare,” International Journal of Engineering Trends and Technology, vol. 71, no. 1, pp. 271–282, Jan. 2023, doi: 10.14445/22315381/ijett-v71i1p224.

[14] V. Yadav, ‘Machine Learning in Managing Healthcare Workforce Shortage: Analyzing how Machine Learning can Optimize Workforce Allocation in Response to Fluctuating Healthcare Demands’, Progress in Medical Sciences, vol. 7, pp. 1–9, 08 2023.

[15] S. Moyo, T. N. Doan, J. A. Yun, and N. Tshuma, “Application of machine learning models in predicting length of stay among healthcare workers in underserved communities in South Africa,” Human Resources for Health, vol. 16, no. 1, p. 68, Dec. 2018, doi: 10.1186/s12960-018-0329-1.

[16] J. Song et al., “The Random Forest model has the best accuracy among the four pressure ulcer prediction models using machine learning algorithms,” Risk Management and Healthcare Policy, vol. Volume 14, pp. 1175–1187, Mar. 2021, doi: 10.2147/rmhp.s297838.

[17] K. Topal, D. Koc, and E. Eren, “Workforce forecasting with machine learning for healthcare management,” Journal of Health Management, 2026, doi: 10.1177/09720634251396339.

[18] B. Köksoy, A. Paşaoğlu, and A. Akbel, “Forecasting regional healthcare workforce demand in Turkey using machine learning algorithms,” 2025.

[19] Seguro Social de Salud (EsSalud), Directiva de Gerencia General N° 012-GG-ESSALUD-2015: Normas de los Procesos de Admisión, Consulta Externa y Atención Ambulatoria en las IPRESS del Seguro Social de Salud – EsSalud, Lima, Peru, 2015. [Online]. Available: https://hdl.handle.net/20.500.12959/754.