ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.902

Predictive and Prescriptive Analytics based on Big Data for the Management of High Blood Pressure

Witman Alvarado-Diaz[1]; Brian Meneses-Claudio[2]
[1,2]Universidad Cientifica Del Sur, Peru, *walvaradod@cientifica.edu.pe*

Abstract - High Blood Pressure (HBP) is one of the leading causes of cardiovascular disease, premature mortality, and catastrophic health expenditure in Latin America, especially when it coexists with diabetes mellitus and progresses to chronic kidney disease (CKD). Despite the widespread availability of effective and low-cost antihypertensive treatments, traditional management models, reactive and focused on isolated clinical episodes, have shown a limited capacity to contain the lack of control and its clinical and financial consequences. That is why our objective is to develop a predictive model of risk of hypertensive decontrol and nephrological progression, and to propose a prescriptive analytics approach to optimize the allocation of financial resources in health. A retrospective study is carried out analyzing 24.2 million transactional records of medical care, pharmacy and diagnoses corresponding to 1.02 million patients treated between 2022 and 2024, using Big Data architecture based on Apache Spark. Different Machine Learning models were trained, and the Gradient Boosted Trees (GBT) model was chosen for the prediction of complications. As a result, we were able to identify a critical data quality gap. The predictive model reached a recall of 89.74%. Prescriptive analysis demonstrated that preventive intervention reduces cost compared to delayed renal treatment. We conclude that the sustainability of health systems requires a transition from reactive models to prescriptive nephroprotection strategies, supported by advanced analytics and systematic auditing of clinical data.

Keywords — Hypertension, Big Data, PySpark, Machine Learning, Prescriptive Analysis, Descriptive Analysis, Public Health.

I. Introduction

High blood pressure (HBP) is the leading modifiable risk factor for cardiovascular disease and premature mortality globally, affecting more than 1.28 billion adults [1]. Despite the availability of effective, safe, and low-cost drug treatments, blood pressure control rates remain insufficient, particularly in low- and middle-income countries, where less than 20% of patients achieve adequate control. [2]

In the Peruvian context, the epidemiological transition towards chronic non-communicable diseases has steadily increased the demand for primary care services. Recent reports from the Demographic and Family Health Survey (ENDES) indicate that, although access to the diagnosis of hypertension has improved, the proportion of patients with effective control has stagnated or even deteriorated in the post-pandemic period [3]. This situation is aggravated when hypertension coexists with type 2 diabetes mellitus, accelerating the progression to chronic kidney disease (CKD), one of the costliest and clinically devastating complications. [4]

CKD represents an event of high fiscal impact for health systems since the treatment of a minority of patients on renal replacement therapy can consume up to 20% of available health budgets [5]. In this scenario, the challenge no longer lies in the availability of widely accessible antihypertensive drugs but in the early detection of lack of control, as well as the efficient allocation of preventive resources.

The increasing digitization of electronic health records (EHRs) has generated massive volumes of clinical and administrative data that exceed traditional analysis capacity. Nonetheless, this "deluge of data" has yet to be fully translated into actionable insights due to analytical and data quality limitations [6]. In this context, the use of Big Data architecture and Machine Learning (ML) algorithms offers an unprecedented opportunity to transform HTN management, allowing the identification of complex and non-linear risk patterns that traditional clinical scores fail to capture. [7][8]

Recent evidence shows that supervised learning models, such as Random Forest and Gradient Boosting, outperform conventional risk scales by integrating multiple comorbidities, polypharmacy schemes and administrative variables in real care settings [9], also showing superior performance compared to linear models, especially when combined with under sampling techniques in order to manage the class imbalance inherent in clinical data, [10] [11]. In the cardiovascular field, Pulse Pressure (PP) has emerged as a superior predictor of arterial stiffness and cardiovascular risk in older adults, surpassing isolated systolic pressure, which has driven its incorporation into advanced risk models. [9]

Studies of importance of variables agree that Diabetes Mellitus is one of the most relevant predictors of cardiovascular and renal complications, displacing traditional demographic factors such as age and sex [12] [13]. However, it has been documented that the quality of clinical data constitutes a significant structural limitation, which restricts epidemiological surveillance and the performance of predictive models. [14][15][16]

From an operational perspective, the evidence supports models of care in which active nursing and nutrition participation improves blood pressure control, aligning with the recommendations of the World Health Organization for resource-limited contexts [4],[17][18]. Finally, the literature suggests that the integration of home monitoring strategies could substantially improve data quality and raise the predictive capacity of models beyond AUC values close to 0.85 [19], [20].

II. Methodology

It is considered of utmost importance, to start by mentioning our source of data, which have been obtained from an open data source owned by the Peruvian state, the data can be located on the website https://www.datosabiertos.gob.pe/, this page is a free service of the Peruvian state, in which different databases of the different sectors of the Peruvian state are made available to interested people; we have obtained the database called "Health benefits associated with Insured with Arterial Hypertension - [SIS]", which has been provided by the comprehensive health insurance (SIS), of Peru. [21], [22]

For this project we will apply as much as possible the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, described extensively in [23], this is a standard methodology for managing data mining and data science projects in a structured, repeatable and business-oriented way. It is an iterative and flexible methodology, suitable for Big Data and health projects as is our case, in the following image we can graphically show the complete cycle of the methodology that will be applied in this work.

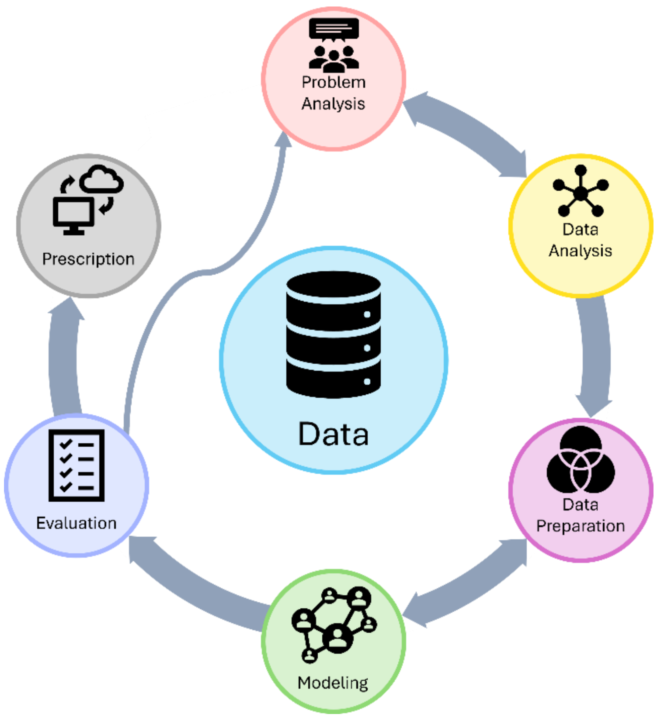

Fig. 1. CRISP-DM Methodology Chart

The CRISP-DM methodology is composed of six phases related to each other, which do not follow a strict linear flow, but iterative; The phases of this methodology are:

A. *Problem Analysis*

This phase has as its main objective to understand the problem and translate them into concrete and measurable analytical objectives [21]. In this stage, the strategic objectives are identified, criteria are defined and the technical, legal, and operational constraints of the project are analyzed. [23]

B. *Data Analysis*

In this phase, the initial data are collected, and an exploration analysis is carried out to understand its structure, content, and quality [21] [24]. In addition, potential quality issues such as incomplete data, duplicates, or noise, which can significantly affect the performance of analytical models, are evaluated. [22]

C. *Data Preparation*

The Data Preparation phase consists of the construction of the final dataset that will be used for modeling, integrating multiple sources of information and applying processes of cleaning, transformation, and selection of attributes [21] [24]. Numerous studies indicate that this stage can consume up to 70% of the total project time, due to the complexity and heterogeneity of the actual data [23].

D. *Modeling*

In the Modeling phase, data mining or machine learning techniques are selected and applied to the prepared dataset [21], [24]. Since different techniques may require different data formats, it is common for this phase to involve feedback towards the data preparation stage. [22]

E. *Evaluation*

The purpose of the Evaluation phase is to verify that the models developed meet not only technical performance criteria, but also with the objectives of the proposed model [21]. In this stage, metrics such as accuracy, sensitivity, specificity, or AUC are analyzed, depending on the type of problem addressed [24]. This phase allows you to decide if your project is ready for deployment or if you need to iterate back to earlier phases. [22]

F. *Prescription*

This consists of the implementation of the results obtained for practical use, either through the integration of models in production systems, the generation of reports or the development of decision-making support tools [21]. CRISP-DM emphasizes that the deployment does not mark the end of the process, but the cycle of continuous monitoring and maintenance of the model, in order to ensure its validity and effectiveness over time.[22], [23]

III. Results

In this stage we will see the results obtained from each stage of the methodology applied, described in the previous section.

A. *Analysis of the problem*

The central problem does not lie in the absence of pharmacological therapies or in the lack of health coverage, but in the inability of the system to identify early patients with a higher risk of sustained lack of control and progression to high-cost complications, such as chronic kidney disease (CKD).

From a clinical perspective, the management of hypertension is based on isolated consultations and non-systematized clinical decisions, ignoring the longitudinal nature of the disease, which limits the detection of patterns of progressive deterioration, especially in patients with metabolic comorbidities such as diabetes mellitus. As a consequence, a considerable proportion of patients move from hypertensive decontrol to chronic renal failure, at which point the available interventions are costly and of limited impact.

Overall, the problem to be solved is not only clinical, but systemic: the absence of an integrated analytical framework that allows auditing data quality, predicting individual risk and prescribing cost-effective interventions prevents the transition to a preventive, sustainable and value-based model.

B. *Data analysis*

The data obtained are made up of 5 CSV files, which are: HTA_Diagnosticos, HTA_FUAS, HTA_Insumos, HTA_Medicamentos and HTA_Procedimientos, from which historical information was extracted from 2022 to 2024, covering a universe of 1,026,473 unique patients and a total of 24,286,707 transactional records (HTA_FUAS). The processing of all this information was done using Apache Spark (PySpark) due to the volume of data.

In the following image we can see different graphs among which we have gender distribution (graph A) where the female gender is the majority class with 66% of the records while the class of men covers 33% of the records, another of the graphs shows the number of visits by level of health facilities (graph B). where we can see that 60% of the population is treated in level 1 health centers, in addition to 20% and 19% for levels 3 and 2, respectively. In graph C we have the distribution of the data corresponding to the age variable where it is observed that most of the population is approximately between 50 and 80 years old. Finally, graph D shows the number of patients by department, with Lima being the department with the highest presence of cases, corresponding to 35% of the records.

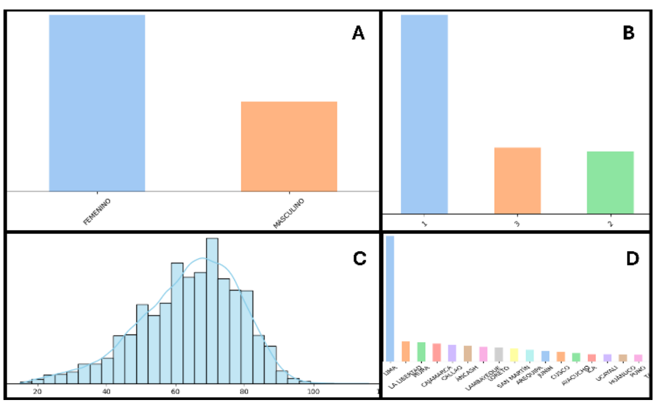

Fig. 2. Graphs for data analysis.

During this stage, a severe gap in the capture of essential clinical information was detected, which compromises the patient's ability to monitor. 49.75% of the care records (almost 12 million rows) do not have a valid Blood Pressure value (they register 0 or NULL), with respect to the clinical status of these we do not know if they are patients who were controlled or in crisis; The most likely cause of such a large number of erroneous records is that there are failures in the registration protocols, whether in triage, nursing or administrative care. Another of the greatest difficulties encountered is that 98.5% of the variable corresponding to the days of hospitalization are null or missing values, making this an irrecoverable variable since it has too much missing information, therefore, it will not be taken into account in the following stages.

TABLE I. TABLE OF COMORBIDITIES

CODDIA	COUNT	NAME
I10x	5118712	Hypertension
E119	1566644	Type 2 diabetes mellitus
N390	1039254	Urinary tract infection
E785	663627	Hyperlipidemia

In the table above we can see an analysis of comorbidities and we see them with their corresponding number of records, which show that the user profile is complex, where it is confirmed that hypertension does not travel alone, since its inseparable companions are: Diabetes mellitus, dyslipidemia or cholesterol and kidney infection, these diseases make up a triad that is common in patients with high cardiovascular risk.

TABLE II. MEDICATION EXPENDITURE TABLE

MEDICATION	FREQUENCY	GASTO_TOTAL
Peritoneal Dialysis Solution	18876	39604722.83
Medical oxygen	153424	37807961.5
Meroponem	42118	14723572.87
Losartan potassium	2585141	9641207.88

In the table above, Losartan Potassium is the drug with the highest volume (2.5 million deliveries), which is logical since it is necessary to treat HTN. However, its cost is low, which is also adequate since patients can access the drug more easily. According to the table we realize that the expense is going into severe complications, since the solution for peritoneal dialysis is used in patients with kidney failure, as well as Oxygen and Meropenem, which are used in hospitalizations and intensive care. As we can see, poorly controlled hypertension is leading to Chronic Kidney Failure and that is where the financial budget skyrockets.

C. *Data preparation and Modeling*

Taking into account that uncontrolled pressure can generate chronic kidney failure, a machine learning model was created to predict whether a patient will have uncontrolled hypertension (systolic pressure > 140 and diastolic pressure > 80) or is controlled.

In the following image we can see the entire process of data preparation, which begins with the definition of the target variable, for this we filter the data of the systolic and diastolic pressure variable and assign them a "1" if it is uncontrolled hypertension (systolic pressure > 140 and diastolic pressure > 80) or "0" for controlled HTN; Next, the Under Sampling method is used to balance both classes, thus obtaining a balanced target ready for the next stages. The data are divided into 2 groups (training and test) applying a proportionality of 70 – 30, respectively. Within our data we have categorical variables, so it is necessary to index and encode them as a vector so that ML algorithms can understand and process them. The data passes through a pipeline that is responsible for guiding the entire preparation process in order to obtain data ready for training with ML models. It should be mentioned that we have tested 5 different ML models among them we have: Logistic Regression, Gradient-Boosted Trees (GBT) Classifier, Random Forest, Multilayer Perceptron (MLP), Linear Support Vector Classification (SVC), from each of these models the corresponding metrics were obtained which will help us evaluate their performance.

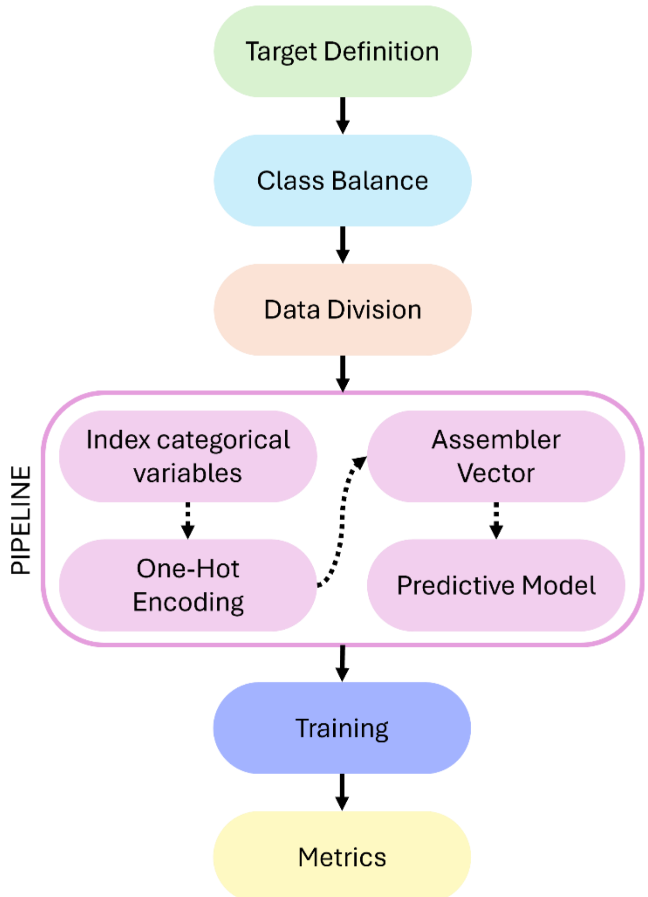

Fig. 3. Diagram of the preparation and modeling of the data.

D. *Evaluation*

For the analysis, we must take into account that predictive models try to detect uncontrolled hypertension in a population that is already hypertensive; In this medical scenario, the cost of a False Negative (telling a patient at risk that they are fine) is much higher than that of a False Positive (alerting a doctor about a patient who happens to be under control); Therefore, we conclude that our star metric will be Recall, obviously without losing sight of Accuracy so as not to saturate the health system with false alarms. Taking this into account, in the following table we can see a summary of the most important model metrics of each of the trained models.

TABLE III. MOST IMPORTANT MODEL METRICS

MODEL	AUC	RECALL (%)	ACCURACY (%)	F1-SCORE	FALSE NEGATIVE RATE (FNR) (%)
GBT Classifier	0.6190	89.74	57.62	0.7018	10.26
Logistic Regression	0.7438	65.24	70.11	0.6759	34.76
Linear SVC	0.7431	59.37	75.44	0.6645	40.63
MLP	0.7502	57.11	76.15	0.6527	42.89
Random Forest	0.7575	52.54	80.92	0.6372	47.46

Logistic Regression: This model presents a behavior in which it does not take too much risk, It has a decent discriminatory capacity (AUC 0.7438), It detects 2 out of 3 patients at risk (Recall 65.2%); However, it allows 35% of severe patients to pass (FNR 0.347), which is a moderate-high clinical risk. It is a robust and explainable model, but for an early warning system, missing 35% of critical cases could be unacceptable.

Random Forest: This model looks for purity in its positive predictions but is too cautious. Mathematically, it is the model that best separates classes (AUC 0.7575), If this model says that someone is at risk, it is very likely to be true (Accuracy 80.9%), its negative part is that it is ignoring almost half of patients with uncontrolled hypertension (FNR 0.47) this model is unfeasible for primary risk detection and prevention.

GBT Classifier: This model has a decision threshold set to 30%, i.e., the model has been forced to be extremely sensitive. This radically changes their behavior compared to others; It detects 9 out of 10 patients at risk (Recall 89.7%). It is the best model to ensure that almost no one is left without care. The cost of high sensitivity is that it classifies many healthy patients as at risk as it has a false positive rate of 66%. The AUC of this model (AUC 0.6190) is low compared to the others, which indicates that its ability to sort cases is weaker, but the threshold adjustment makes it more useful for this project objective. This model is ideal for a first stage of triage where it is vital not to ignore anyone, as long as you have personnel to review all false alarms.

Multilayer Perceptron: It behaves very similarly to Random Forest and Logistic Regression, being located in a midpoint. It does not provide a clear advantage over Logistic Regression, nor does it exceed the precision of Random Forest.

Linear SVC: The results are almost identical to Logistic Regression and MLP. It's a competent model, but one that doesn't excel in any critical metrics for this specific problem.

E. *Prescription*

Being an experimental work without the possibility of real intervention, the approach of the Prescriptive Analysis must be strategically simulated oriented to the design of policies and impact estimation, being our objective to reduce uncontrolled hypertension through the implementation of strategies based on the GBT model, for this a simulation "What-If" is carried out, since it is the most powerful tool in an experimental environment. using the GBT model to simulate hypothetical scenarios, for this we propose 3 scenarios:

TABLE IV. PERCENTAGES OF PATIENTS RESCUED

SCENARIO	RESCUED %
No Obesity	0.72
Pulse Controlled Pressure	3.78
Maximum 3 Drugs	0.04
Multidisciplinary	4.75

The fact that the percentages are low suggests that patients have multiple simultaneous comorbidities; If we remove only obesity from an obese, diabetic patient with high pulse pressure, the model says: "They are still diabetic and have high blood pressure, so the risk remains."

Finally, we have the results of a multidisciplinary action, which shows us a total reduction of 4.75%, which at first glance might seem discouraging, but analytically it is a forceful conclusion since it represents a rescue of more than 120,000 patients.

A critical finding is the remaining 95% of the population that is still identified in the Risk category even after optimizing all modifiable clinical variables. This indicates that uncontrolled hypertension in the population analyzed is structural since we cannot prevent it with standard health campaigns, so we must prepare to manage the chronicity of the disease.

Since we found that 95% of the risk is structural, the analysis must now focus on how to efficiently manage that large block of patients, for which an economic impact analysis was carried out. Where it is obtained that the services of prevention and control types on average have a cost of care of S/ 15.19, while the care services with serious diagnoses have an average cost of care of S/ 1699.79, which reflects that the cost of making a mistake in prevention is irrelevant compared to the savings of avoiding a single serious event.

By counting the false negatives, false positives, and true positives of the percentage with structural risk, we obtain the following table.

TABLE V. PATIENT COUNT BY ECONOMIC SCENARIO

ECONOMIC SCENARIO	COUNT	DESCRIPTION
Undetected cost	158328	Model error (False negative) we did nothing and the patient was sick
Natural savings	512811	Negative hit, no one cat anything
Extra preventive expenditure	1087308	False alarm (false positive) unnecessary spending on prevention
Effective investment	1440536	Success (true positive) spending on prevention, we save serious event.

According to the table:

Investment in prevention

- Total, patients to be operated on 1440536(TP)+ 1087308(FP) = 2527844 patients.
- Preventive Unit Cost: 15.19
- Total Operating Cost: 2527844 * 15.19 = 38372950.36 (38.3 million)

Patients who were really going to have a serious event (PT), but that the model detected in time.

- Patients Rescued: 1,440,536
- Event Cost Avoided: 1,699.79
- Projected Gross Savings:1440536 * 1,699.79 = 2,448,608,687 (2,448 million)

Return on Investment (ROI)

$$ROI = \frac{Total\ Revenue - Total\ cost}{Total\ cost}$$

$$ROI = \frac{2448\ M - 38.3\ M}{38.3\ M} = 62.9164$$

The implementation of the preventive strategy on the 2.5 million patients detected by the model would require an operational investment of 38.3 million. Of these, by timely detecting 1.44 million patients of real elevated risk, a potential expenditure on emergency services and hospitalization estimated at 2448 million soles is avoided. So, for every coin invested in this prevention program, the institution saves 63 soles in future medical expenses.

We must take into account that not all patients of that 95% structural risk are the same, some are renal, others are frail elderly, others are complex diabetics, because of this the K-Means clustering algorithm was executed, only on the patients that the model predicted as High Risk, in order to find some similarity profiles among the patients in this group.

TABLE VI. CLUSTERS FOUND

ID	TOTAL	AGE	PP	COST	RENAL RATE	DIABETES RATE
0	1058419	74	49	17.97	0.0113	0.1102
1	686479	69	71	51.40	0.0256	0.1324
2	782946	52	42	50.99	0.0207	0.1364

With the K-Means algorithm, 3 noticeably clear patient profiles were found, which we can see in the table above, which indicates that we cannot treat all patients in the same way.

Cluster 0: is the largest group (1.05m). Advanced age (74 years), but normal pulse pressure (49), low cost, and low renal rate. This group is made up of elderly people with natural hypertension due to age, but stable and cheap to maintain. This group can be provided with home pharmacy, telemonitoring, nursing. This group does not require seeing the specialist in person.

Cluster 1: more than 686 thousand patients, medium-high age (69), extremely high pulse pressure 71.7, higher renal rate (2.5%) and prohibitive cost (51.4). In this group, pulse pressure > 60 indicates rigid arteries and imminent risk to health, it is the most expensive and critical group, immediate review by cardiology is recommended, evaluate the possibility of some therapy to lower arterial stiffness, in addition they must have urgent renal protection.

Cluster 2: more than 782 thousand patients, with a normal pulse pressure, but with a high rate of diabetes, this is the youngest group, but with a similar expense as the elderly due to diabetes, with this group you should invest in nutrition and endocrinology, with the aim of controlling diabetes and weight as soon as possible to prevent this group from becoming cluster 1 in the future.

We must also take into account that the GBT model returns a probability, so a patient with a 95% probability is very different from one with 34%, even though both are classified as patients at risk; For this reason, priority levels based on gross probability were created, which are as follows:

- Red Zone (Probability > 0.80): Immediate action.
- Orange Zone (Probability between 0.50 and 0.79): Short-term action.
- Yellow Zone (Probability between 0.30 and 0.49): Periodic monitoring.

Having the established levels we must know in which regions of Peru we must act first, for this we cross the information based on the departments and by health establishment and only the red zone that is the most urgent, with this we obtain the following table.

TABLE VII. OPERATIONAL GAPS

DEPARTMENT	EESS LEVEL	COUNT
LIMA	1	82300
LORETO	1	32144
FREEDOM	1	30810
PIURA	1	29926
CAJAMARCA	1	26624

Here we detected a serious failure in the allocation of resources, since there are patients classified as critical (red) being treated in medical posts (level 1), which do not have ICU or adequate specialists for their care. The Lima case (82,300 critical patients in level 1) is a problem of saturation, hospitals are full and serious patients bounce back to neighborhood posts. It is recommended to create rapid referral centers, especially in the sectors furthest from Lima, exclusively for patients with a risk probability greater than 80%. The case of Loreto (32,144 critical patients at level 1) is a problem of access since Loreto has a difficult jungle geography. 32,000 people in serious condition in rural posts are a time bomb. Assisted telemedicine must be implemented in order to provide the Loreto posts with telemetry equipment so that a specialist in Lima or another region of Peru can adjust the treatment remotely, because it is impossible to transfer 32,000 people, it is necessary to prioritize the transfer of knowledge.

IV. Discussions

The results confirm that the management of hypertension cannot be dissociated from the management of diabetes. The high correlation between both pathologies and the risk of kidney damage underscores the need for integrated cardio-metabolic clinics rather than separate vertical programs. [25][26]

The high amount of missing data, close to 50%, is an alarming finding that coincides with the literature on the fatigue of recording in electronic medical records [16]. This limits the ability of algorithms to detect seizures in real-time, suggesting the need to automate the capture of vital signs using IoT or wearables [19][17][18]

The main limitation of this study is the absence of data on lifestyle habits and actual pharmacological adherence, variables that could increase the accuracy of the model above 0.85 AUC as other models in .[20]

Our findings suggest that the management of hypertension in Peru should pivot from an age-focused approach to one focused on metabolic comorbidity. The strong predictive association between diabetes and hypertension reinforces the need for integrated clinical practice guidelines, as recommended by the most recent international standards.[26][16]

In this work, no studies were carried out by medical specialties, however, this is recommended as a cost-effective strategy endorsed by the WHO for health systems with limited resources [19] and [4]. In addition, it is supported by other studies where it is mentioned that doctors should be reserved for complex cases, while routine monitoring should be led by nursing and nutrition, who demonstrate greater adherence to follow-up protocols, [25] and [26] this should be taken into account when proposing intervention policies.

In the studies [12], [13] and mention the current challenges in the management of hypertension. It is mentioned that the predictive importance of diabetes (84.5%) confirms that hypertension in the 21st century is fundamentally a cardio-metabolic disease. This confirms the finding in our study that patients have metabolic diseases associated with hypertension.

This study is based on administrative data, so it lacks variables on lifestyle habits that could improve prediction. Likewise, pharmacological adherence is an indirect estimate based on medication collection, but not on actual intake.

V. Conclusions

The application of advanced analytics on healthcare Big Data made it possible to comprehensively redefine the Arterial Hypertension (HTN) management strategy in the health network studied. The results show that the clinical and financial sustainability of the program depends, predominantly, on the prevention of chronic kidney disease in patients with diabetes mellitus, confirming the significant role of metabolic risk as a driver of hypertensive decontrol.

From a predictive point of view, the study demonstrates that the selection of the optimal model should not be based exclusively on traditional global metrics such as AUC or accuracy. Although the Random Forest model presented superior technical performance in terms of AUC, its low sensitivity limited its clinical utility by not identifying a substantial proportion of patients at risk. In contrast, the Gradient Boosted Trees model adjusted to maximize sensitivity showed superior operational performance, reaching a detection capacity close to 90% of cases of lack of control, which makes it the most suitable alternative for a prevention-oriented public health environment.

The financial prescriptive analysis confirmed that the low accuracy associated with a higher false positive rate does not compromise the economic viability of the program, given the low unit cost of preventive interventions compared to the prohibitive cost of advanced renal therapy. In this context, it is financially rational to prioritize models with high sensitivity, minimizing the risk of false negatives, whose economic and clinical impact is substantially greater.

Overall, this study demonstrates that the transition from a reactive and curative approach to a preventive predictive model, based on advanced analytics and financial prescription, is not only technically feasible, but imperative for the sustainability of health systems in resource-limited contexts.

VI. RECOMMENDATIONS

It is proposed to adopt prescriptive models based on metabolic risk to prioritize preventive interventions in patients with hypertension and diabetes, maximizing the health impact and efficiency of spending by reducing progression to chronic kidney disease. Programs should evolve towards an early nephroprotection approach, systematically integrating renal and metabolic monitoring.

Likewise, it is key to implement automated data quality audits and advance in the automation of the capture of vital signs to reduce underreporting and improve real-time surveillance. Operationally, predictive models with high clinical sensitivity should be prioritized and the use of advanced analytics in budget planning and institutional management should be formalized.

REFERENCES

[1] K. T. Mills, A. Stefanescu, and J. He, "The global epidemiology of hypertension," *Nat Rev Nephrol*, vol. 16, no. 4, pp. 223–237, Apr. 2020, doi: 10.1038/S41581-019-0244-2.

[2] D. Villarreal-Zegarra, R. M. Carrillo-Larco, and A. Bernabe-Ortiz, "Short-term trends in the prevalence, awareness, treatment, and control of arterial hypertension in Peru," *Journal of Human Hypertension 2020 35:5*, vol. 35, no. 5, pp. 462–471, Jun. 2020, doi: 10.1038/s41371-020-0361-1.

[3] "KDIGO 2024 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease," *Official Journal Of The International Society Of Nephrology*, vol. 105, no. 45, Apr. 2024, doi: 10.1016/S0085-2538(24)00110-8.

[4] C. Park, G. Wang, J. M. Durthaler, and J. Fang, "Cost-effectiveness Analyses of Antihypertensive Medicines: A Systematic Review," *Am J Prev Med*, vol. 53, no. 6 Suppl 2, p. S131, Dec. 2017, doi: 10.1016/J.AMEPRE.2017.06.020.

[5] S. Imran, T. Mahmood, A. Morshed, and T. Sellis, "Big data analytics in healthcare. A systematic literature review and roadmap for practical implementation," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 1, pp. 1–22, Jan. 2021, doi: 10.1109/JAS.2020.1003384.

[6] E. Septian *et al.*, "Prediction of Personalised Hypertension Using Machine Learning in Indonesian Population," *J Med Syst*, vol. 49, no. 1, Dec. 2025, doi: 10.1007/S10916-025-02253-5.

[7] G. F. S. Silva, T. P. Fagundes, B. C. Teixeira, and A. D. P. Chiavegatto Filho, "Machine Learning for Hypertension Prediction: a Systematic Review," *Curr Hypertens Rep*, vol. 24, no. 11, pp. 523–533, Nov. 2022, doi: 10.1007/S11906-022-01212-6.

[8] M. Armbrust *et al.*, "Delta Lake: High-Performance ACID Table Storage over Cloud Object Stores," *Proceedings of the VLDB Endowment*, Vol. 13, 2020, doi: 10.14778/3415478.3415560.

[9] H. Jin, S. Fang, S. An, and Y. Ding, "Association of Pulse Pressure Index With Mortality in Patients With Hypertension: Results From NHANES 1999–2018," *The Journal of Clinical Hypertension*, vol. 27, no. 1, p. e70004, Jan. 2025, doi: 10.1111/JCH.70004.

[10] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep Neural Networks and Tabular Data: A Survey," *IEEE Trans Neural Netw Learn Syst*, vol. 35, no. 6, pp. 7499–7519, Jun. 2022, doi: 10.1109/TNNLS.2022.3229161.

[11] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artif Intell Rev*, vol. 57, no. 10, Oct. 2024, doi: 10.1007/S10462-024-10884-2.

[12] "10 Cardiovascular Disease and Risk Management: Standards of Care in Diabetes-2024," *Diabetes Care*, vol. 47, no. Suppl 1, pp. S179–S218, Jan. 2024, doi: 10.2337/DC24-S010.

[13] World Health Organization, "Guideline for The pharmacological tretment of hypertension in addults," *Sustainability (Switzerland)*, vol. 11, no. 1, pp. 1–48, 2021.

[14] R. M. Carrillo-Larco, W. C. Guzman-Vilca, and A. Bernabe-Ortiz, "Mean blood pressure according to the hypertension care cascade: Analysis of six national health surveys in Peru," *The Lancet Regional Health - Americas*, vol. 1, p. 100016, Sep. 2021, doi: 10.1016/j.lana.2021.100016.

[15] S. C. Yeo, H. Wang, Y. G. Ang, C. K. Lim, and X. Y. Ooi, "Cost-effectiveness of screening for chronic kidney disease in the general adult population: a systematic review," *Clin Kidney J*, vol. 17, no. 1, Jan. 2023, doi: 10.1093/CKJ/SFAD137.

[16] S. Omboni and R. Ferrari, "The role of telemedicine in hypertension management: focus on blood pressure telemonitoring," *Curr Hypertens Rep*, vol. 17, no. 4, pp. 1–13, Apr. 2015, doi: 10.1007/S11906-015-0535-3.

[17] G. S. Stergiou *et al.*, "European Society of Hypertension practice guidelines for office and out-of-office blood pressure measurement," *J Hypertens*, vol. 39, no. 7, pp. 1293–1302, Jul. 2021, doi: 10.1097/HJH.0000000000002843.

[18] T. Unger *et al.*, "International Society of Hypertension Global Hypertension Practice Guidelines," *Hypertension*, vol. 75, no. 6, pp. 1334–1357, Jun. 2020.

[19] eBioMedicine, "Machine learning in hypertension: from diagnosis to personalised medicine," *EBioMedicine*, vol. 92, p. 104658, Jun. 2023, doi: 10.1016/J.EBIOM.2023.104658.

[20] S. Kuo, W. Ye, D. Wang, L. N. McEwen, C. V. Santos, and W. H. Herman, "Cost-effectiveness of the National Diabetes Prevention Program: A Real-world, 2-Year Prospective Study," *Diabetes Care*, vol. 48, no. 7, p. 1180, Jul. 2024, doi: 10.2337/DC24-1110.

[21] P. Chapman *et al.*, "CRISP-DM 1.0 Step-by-step data mining guide," 1999.

[22] R. Wirth and J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining".

[23] C. Shearer, "The CRISP-DM Model: The New Blueprint for Data Mining," *Journal Of Data Warehousing*, vol. 5, 2000.

[24] I. H. Witten, E. Frank, and M. A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques, Third Edition," *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*, pp. 1–629, Jan. 2011, doi: 10.1016/C2009-0-19715-5.

[25] World Health Organization (WHO), "Global report on hypertension: the race against a silent killer," *World Health Organization*, pp. 1–291, 2023.

[26] G. Ogedegbe *et al.*, "Task shifting interventions for cardiovascular risk reduction in low-income and middle-income countries: a systematic review of randomised controlled trials," *BMJ Open*, vol. 4, no. 10, p. e005983, Oct. 2014.