International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Improving Accuracy: Comparative Analysis of Machine Learning Models for Prostate Cancer Prediction

Saul Beltozar-Clemente*[1], Enrique Diaz-Vega[2], Isaac Conde Ramos[3], Raul Tejeda Navarrete[4]

Submitted: 18/09/2023 **Revised**: 18/11/2023 **Accepted**: 29/11/2023

Abstract: Among the different types of cancer affecting men is prostate cancer, which ranks second in mortality after lung cancer, a worrying reality. Nowadays, Machine Learning (ML) models have contributed to different areas, being their contribution to the medical field one of the most outstanding. This study aims to compare the accuracy of ML models in the prediction of prostate cancer. Gradient Boosting (GB), Random Forest (RF), Decision Tree (DT) and Adaptive Boosting (AdaBoost) models were analyzed. In addition, DT, RF, K-Nearest Neighbors (KNN), Gaussian Naive Bayes (GNB) and Logistic regression (LR) models were used to identify the base model for algorithm optimization. The study was divided into several stages, such as the description of the models and the analysis of the data set, among others. On the other hand, the metrics of sensitivity, precision, specificity, accuracy, and F1 count were used to contrast the algorithms. The training results positioned the GB algorithm as the most accurate algorithm for prostate cancer detection with 83.33% accuracy, 98.02% precision and 95.24% sensitivity.

Keywords: *Accuracy, comparative, machine learning, prostate cancer*

1. Introduction

According to the latest update of the GLOBOCAN 2020 database, cancer is the leading cause of death worldwide[1], with 20 million new cases diagnosed and 10 million deaths attributed to this disease in its different types[2]. The global cancer burden is expected to increase to 30 million new cases by 2040, with the greatest growth in low- and middle-income countries[3], which ranks second in mortality after lung cancer [4], causing the death of 375,304 people worldwide [5]. Prostate cancer is a malignant disease that originates in the prostate, a gland of the male reproductive system[6], it develops when prostate cells begin to grow abnormally and uncontrolled, forming tumors[7] as the cancer progresses it can invade surrounding tissues and in advanced cases can metastasize[8][9]. Although the exact causes of prostate cancer are not known with certainty[10], several risk factors have been identified that may increase the likelihood of this disease[11][12][13] such as age[14], family history[15][16], race[17][18], BPH[19][20], exposure to chemical agents[21][22] and obesity[23] [24].

In 2020, 1,414,259 people were diagnosed with this disease, representing 7.3% of the total cases. The continents where these cases are distributed are Europe 33.5%, Asia 26.25%, North America 16.9%, Latin America and the Caribbean 15.2%, Africa 6.6%, and Oceania with 1.6%. The mortality rate is 26.5%, which makes 375,304 deaths from prostate cancer, distributed in Asia 32.1%, Europe 28.8%, Latin America and the Caribbean 15.3%, Africa 12.6%, North America 9.95%, and Oceania with 1.3% [25][26].

ML models, has radically transformed the way big data analysis is performed in various industries and disciplines ML models, has radically transformed the way big data analysis is performed in various industries and disciplines[27]. This makes ML a powerful tool for clinical data analysis in healthcare and biomedical research[28]. ML algorithms can learn from clinical data sets to predict medical diagnoses[29], identify subtle patterns in clinical data that might indicate the early presence of disease[30]. Natural language processing in ML is used to analyze electronic medical records and clinical notes, which facilitates the extraction of relevant information[31][32]. It is important to note that the use of ML in clinical data should be approached with caution and comply with privacy and data security standards [33].

2. Related Work

The following are some studies exploring the prediction of prostate cancer using ML models. In[34], the value of machine learning (ML) models in predicting the Ki67 index and GGG of CaP was investigated. A total of 122 patients who had undergone preoperative MRI were included. Logistic regression (LR), support vector machine (SVM), random forest (RF) and K-nearest neighbor (KNN) models were constructed. The results they found indicate that the model (LR_ADC+T2, AUC=0.8882) performed best in Ki67 prediction, and (SVM_DWI+T2, AUC=0.9248)

[1] *Dirección de cursos básicos, Universidad Científica del Sur, Lima, Perú*
ORCID ID : 0000-0002-3742-6326
[2] *Departamento de Ciencias, Universidad Privada del Norte, Lima, Perú*
ORCID ID : 0000-0003-1886-0693
[3] *Facultad de Ingeniería, Universidad Tecnológica del Perú, Lima, Perú*
ORCID ID : 0009-0009-5861-2564
[4] *Departamento de Ciencias, Universidad Tecnológica del Perú, Lima, Perú*
ORCID ID : 0000-0002-3301-9918
* *Corresponding Author Email: sbeltozar@cientifica.edu.pe*

performed best in GGG prediction. Also, in [35] developed a predictive model to improve the accuracy of prostate cancer (CaP) in patients with PSA≤20 ng/ml, 146 patient were evaluated, significant predictors ($p<0.05$) were included in five machine learning algorithm models. A decision curve analysis (DCA) was performed to estimate the clinical utility of the models. Cross-validation was applied ten times in the training process. The results showed that the Random Forest model exhibited the best predictive performance and had the highest net benefit compared to the other algorithms, with an area under the curve of 0.871. In addition, DCA had the highest net benefit across the range of cutoff points examined.

Similarly,[35] , they proposed to develop and validate a machine learning model to identify P504/P63 status and achieve better CaP diagnosis. This study used T2WI, DWI and ADC sequences to evaluate prostate diseases, the P504s/P63 prediction models, P504s/P63 were established using random forest (RF), gradient boost decision tree (GBDT), logistic regression (LR), adaptive boost (AdaBoost) and K-nearest neighbor (KNN) algorithms. The results show that the RF algorithm performed the best in overall evaluations (micro-average AUC = 0.920, macro-average AUC = 0.870) and provided the most accurate result in additional sub-label predictions, the accuracies of labels 0, 1 and 2 were 0.831, 0.831, and 0.932, respectively. In the study [36] performed a systematic evaluation of 15 machine learning (ML) algorithms and 30 gene expression-based prognostic signatures of 1558 primary prostate cancer patients from public data repositories. This study showed that survival analysis models outperformed binary classification models for risk assessment and performance of survival analysis methods: regularized Cox model with ridge penalty (Cox-Ridge) and partial least squares (PLS) regression for Cox model (Cox-PLS).

Also, in[37] they built and performed a cross-validation of a machine learning model based on radiomic features of (T_2 WI)-weighted images of PI-RADS 3 lesions to identify clinically significant prostate cancer (csPCa), A total of 240 patients were included, training cohort, n = 188, age range 43 to 82 years; test cohort, n = 52, age range 41 to 79 years. The results show that the trained random forest classifier constructed from the radiomics (T_2 WI) has a good and statistically significant area under the curves (AUC) of 0.76 (p= 0.022) for the prediction of csPCa in the test set. Prostate volume and PSA density showed moderate and non-significant performance (AUC 0.62, P = 0.275 and 0.61, P = 0.348, respectively) for the prediction of csPCa in the test set. Also, in [38] they evaluated and compared the performance of different machine learning models for diagnosing breast cancer.

The study shows that they have applied logistic regression (LR), K-nearest neighbors (KNN), extreme gradient boosting (XGB), gradient boosting (GB), random forest (RF), multilayer perceptron (MLP) and support vector machine (SVM) algorithms, obtaining as results a maximum accuracy of 90.68% for RF compared to LASSO. Similarly, the recall in KNN was 98.80%, the accuracy in MLP was 92.50% and the F1 score in RF was 94.60%

3. Methodology

This section details the research methodology divided into 2 parts. The first part is dedicated to explaining the ML models (LR, RF, DT, GNB, KNN, AdaBoost and GB), the second part explains the development of case studies.

3.1 Description of ML models

Logistic Regression (LR): s a machine learning technique used in binary and multiclass classification. It combines a linear model with the logistic function to predict probabilities. It is trained by cross-entropy minimization and fits training data to estimate class probabilities. Its versatility makes it applicable in a variety of domains, such as spam detection, medical diagnosis, and product classification. Regularization can improve its performance and prevent overfitting. The interpretability and efficiency of LR make it a valuable tool in classification problems. The general equation of the LR model is expressed as the following expression:

$$P\left(Y-\frac{1}{X}\right)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n)}}$$

The equation allows us to calculate the probability that an observation belongs to one of the two classes in a binary classification problem.

Decision Tree (DT): are models that partition the feature space by clear, hierarchical rules. They are appreciated for their interpretability, ability to handle diverse features and resistance to outliers. They find applications in medical diagnostics, customer segmentation, fraud detection and price prediction. DT models are valuable tools for classification and regression problems due to their simplicity and versatility, facilitating automated decision making based on conditional rules derived from a tree constructed during training.

Random forest (RF): This model combines multiple decision trees to improve accuracy and generalizability in classification and regression problems. In its construction, random sampling with replacement is used to generate Bootstrap sets and individual decision trees are created in each of the Bootstrap sets. The predictions from these trees are then combined to obtain the final RF prediction. Its advantages include high accuracy, resistance to overfitting, and the ability to handle diverse eatures. It is applied in spam detection, medical diagnosis, and price prediction,

becoming a valuable tool in the ML field. The model architecture is detailed in Fig. 1.

Fig. 1. RF model architecture

Gaussian Naive Bayes (GNB): The GNB model is an extension of the Naive Bayes algorithm that is mainly used in classification. Its operation is based on calculating the conditional probabilities that an observation belongs to a given class, considering the distribution of continuous features. GNB is efficient in training and prediction suitable for data sets with continuous characteristics and is used in pattern recognition, medical diagnosis, sentiment analysis and spam detection when dealing with data following a Gaussian distribution. The general equation for calculating the conditional probability of this model is:

$$P(C_k|x_1, x_2, \dots, x_n) = \frac{P(x_1|C_k)P(x_2|C_k) \dots P(x_n|C_k)P(C_K)}{P(x_1)P(x_2) \dots P(x_n)}$$

This equation is based on Bayes theorem.

K-nearest neighbors: The K-NN model is a supervised learning technique used for both classification and regression problems. In classification, K-NN finds the K data points closest to an unknown point based on some distance metric and assigns the most common class among these neighbors to the unknown point. In regression, it calculates the average of the target values of the K nearest neighbors. It is used in recommendation and pattern recognition systems, medical diagnosis, and social network analysis. The equation detailing this model is:

$$d(x_q, x_i) = \sqrt{\sum_{j=1}^{D} \left(x_q^j - x_i^j\right)^2}$$

The equation in this model is based on the idea of computing the distance between data points to determine their similarity.

Adaptive Boosting: The AdaBoost model is an algorithm for improving the accuracy of classification models. AdaBoost iteratively trains weak classifiers by adjusting the weights of the examples based on errors made by previous classifiers. It then combines these weak classifiers into a strong classifier by weighted voting. The ability to significantly improve accuracy makes it valuable in a wide range of applications such as object detection in imaging as well as medical diagnostics. The general mathematical expression of this model can be represented as follows:

$$F(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$$

Where, $F(x)$ is the strong classification function that is used to make predictions.

Gradient Boosting: this GB technique stands out for its ability to improve the accuracy of regression and classification models. The process focuses on minimizing the residual error in each iteration, which makes it effective for capturing complex relationships in heterogeneous and nonlinear data. Its applications are varied, ranging from weather forecasting, financial analysis, fraud detection and medical diagnosis. Fig. 2 provides a broader view of the model's architecture.

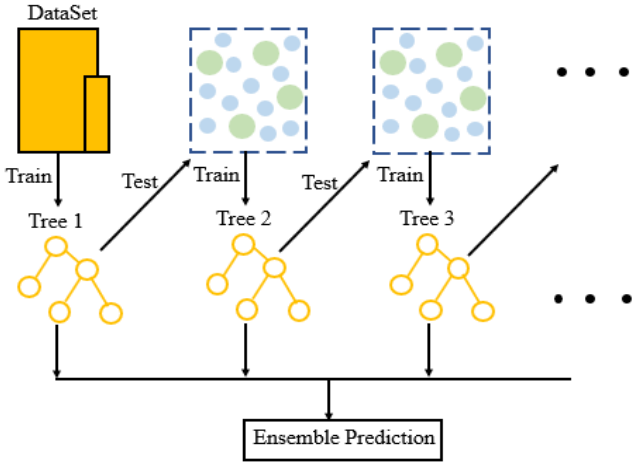

Fig. 2. GB model architecture

3.2 Case Studies

Understanding the data set: The source of the data used in this study was Kaggle, which comprises the medical records of 734 patients and consists of a total of 27 attributes, such as age, number of sexual partners, age at first intercourse, use of condoms during sex (yes=1, no=0), smoking (yes=1, no=0), years of smoking, number of packs of cigarettes smoked per year. have had sex with men, women or both, do physical activities regularly(yes = 1, no = 0), drink alcohol(yes = 1, no = 0), Has any first-degree relative had prostate cancer(yes = 1, no = 0), has any other family member had cancer(yes = 1, no = 0), family history of other genetic or hereditary diseases(yes = 1, no = 0), experience difficulty urinating(yes = 1, no = 0), have had recurrent urinary tract infections(yes = 1, no = 0), have had previous prostate problems such as BPH(yes = 1, no = 0), have had previous prostate biopsy(yes = 1, no = 0), experience pelvic pain(yes = 1, no = 0), have erectile difficulties(yes = 1, no = 0), experience painful ejaculation(yes = 1, no = 0), have noticed changes in urinary pattern recently(yes = 1, no = 0), presence of sexually transmitted diseases(STD)(yes = 1, no = 0). For the attributes on the presence of STDs according to heir ype, it is considered in the same way where "yes" is represented as 1 and "no" is represented as 0: syphilis, genital herpes, AIDS, HIV, hepatitis. The case study development process is presented in Fig. 3.

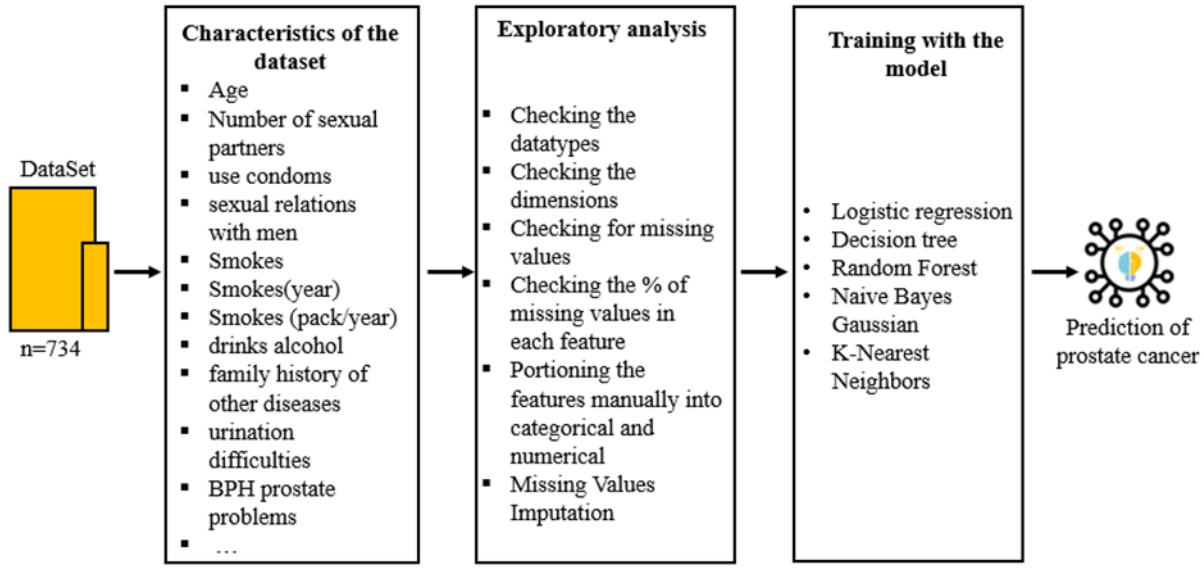

Fig. 3. Case study development process

Data preparation: To start the data processing, we proceeded to import the libraries to perform a general analysis. During this analysis, the content of the 27 variables was explored, missing values were identified, and it was observed that the variable names contained unnecessary spaces. o resolve this, it was decided to add the underscore character (_) to the variable names to eliminate spaces. Regarding the processing of missing values, it was found that there were numerous null values in some variables, as is the case of the variables recording the time of initial and final diagnosis of sexually transmitted diseases, which presented more than 70% of the null. Due to the sensitivity

of the data, it was decided not to use imputation methods such as median or mode. Instead, we chose to use ML models to fill in the missing data. Independent imputation was performed in two groups: those that required synthetic data generation were referred to as sample 'Y', while those that did not require synthetic data were referred to as sample 'X'. Next, the DecisionTreeRegressor model was used for the numerical columns and the DecisionTreeClassifier model for the categorical columns. The models were built and trained, and predictions were made for the sample 'Y' values. The results of the synthetic data generation are presented in Table 1. Finally, we made sure that there are no missing values and checked the statistics of some of the columns, as shown in Table 2.

Table 1. Analysis of the data set

	1	2	3	4	5	…	729	730	732	733	734
Age	18	15	52	46	42	...	34	32	25	33	29
No_of_sex_partner	4	1	5	3	3	...	3	2	2	2	2
First_sexual_intercourse	15	14	16	21	23	...	18	19	17	24	20
Use_condoms	1	1	0	0	0	...	0	1	0	0	1
Smoke	0	0	1	0	0	...	0	0	0	0	0
Smokes_years	0	0	37	0	0	...	0	0	0	0	0
Smokes_packs_year	0	0	37	0	0	...	0	0	0	0	0
sexual_relations_men	0	0	1	1	0	...	0	1	1	1	1
Physical_activities	1	0	1	1	0	...	0	1	0	0	0
Drinks_alcohol	0	1	1	0	0	...	0	0	1	0	0
Family_prostate_cancer	0	0	0	0	0	...	0	0	0	0	0
Other_relative_cancer	0	0	0	1	0	...	0	1	0	0	0
Family_genetic_diseases	0	0	0	0	1	...	0	0	0	1	0
Urination_difficulties	0	1	0	1	0	...	1	0	1	0	0
Recurrent_urinary_tract_infections	0	0	0	0	0	...	0	1	0	0	0
HPB	1	0	0	0	0	...	0	0	1	0	0
Prostate_biopsy	0	0	0	1	0	...	0	0	0	0	0
Pelvic_pain	0	1	0	0	0	...	0	0	0	0	1
Erection_difficulties	0	0	0	1	1	...	1	0	0	0	0
Painful_ejaculation	1	1	0	0	0	...	0	1	1	0	0
Urinary_changes	0	0	1	1	1	...	0	0	0	1	1
STDs	0	0	0	0	0	...	0	0	0	0	0
STD_syphilis	0	0	0	0	0	...	0	0	0	0	0
STD_genital_herpes	0	0	0	0	0	...	0	0	0	0	0
STD_AIDS	0	0	0	0	0	...	0	0	0	0	0
STD_HIV	0	0	0	0	0	...	0	0	0	0	0
STD_Hepatitis	0	0	1	0	0	...	0	0	0	0	0

Table 2. Statistics

	Age	Age_at_first_sexual_intercourse	No_of_sex_partner	STDs_No_of_diagnosis	STDs_number
count	734	734	734	734	734
mean	37.812649	16.278043	2.511337	0.084726	0.15155
std	8.529209	7.450174	1.58967	0.295293	0.52164
min	16	13	1	0	0
25%	29	16	2	0	0
50%	44	15	2	0	0
75%	58	20	3	0	0
max.	88	23	28	3	4

Data exploration: in the univariate analysis of the biopsy column, we observed an imbalance in the diagnoses, only 8.6% of the patients have a positive diagnosis of prostate cancer, this imbalance is significant and should be considered in the ML models. In addition, 88.2% of the patients have not had any sexually transmitted disease. Similarly, there are 2.8% of patients with a positive diagnosis of HIV infection and 8.4% have suffered from BPH as shown in Fig. 4.

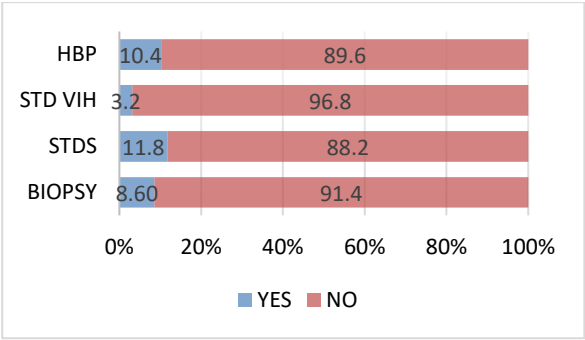

Fig. 4. Statistics

It can be seen from Fig. 5 that there is a higher density of patients between 18 and 40 years of age, most of whom had their first sexual intercourse between 14 and 19 years of age.

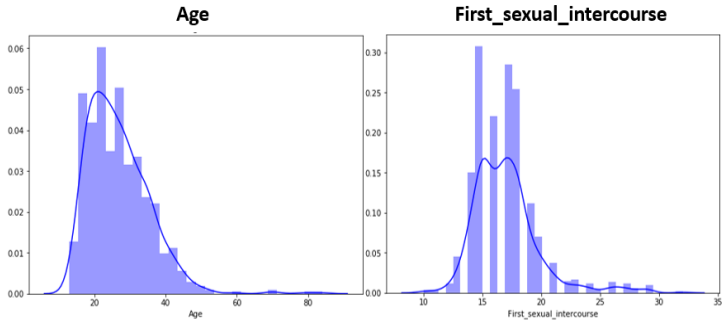

Fig. 5. patient density

Similarly, Fig. 6 shows that prostate cancer can affect both smokers and non-smokers, mainly patients over 22 years of age. In addition, patients who have been smoking for more than a year are more likely to be biopsy positive. It also shows that older patients who smoke large packs of cigarettes per year have a higher susceptibility to be diagnosed with prostate cancer.

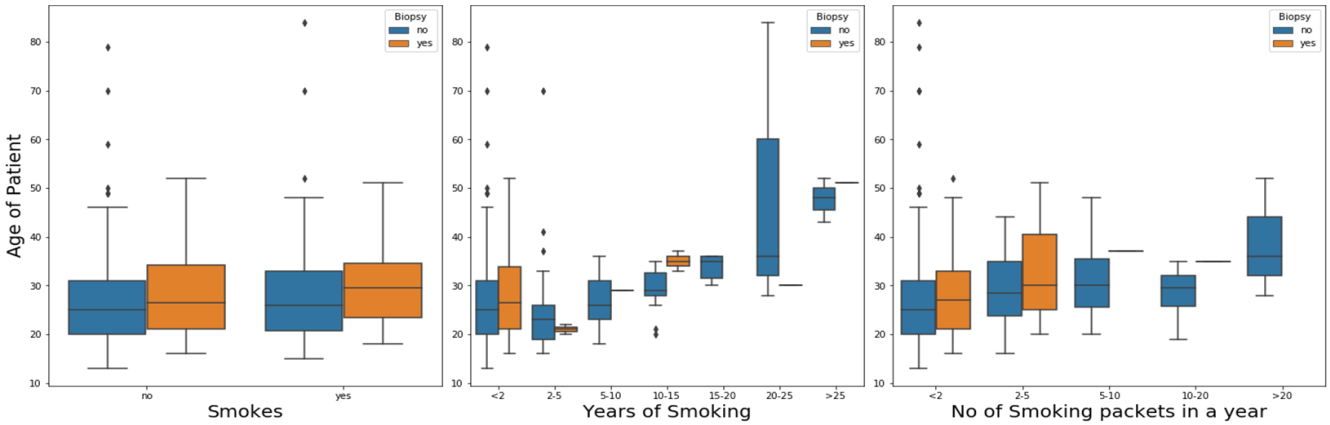

Fig. 6. Analysis of variables age, smoking, years smoked, and number of packs smoked during the year

Data processing and modeling: To ensure that the models achieve the highest possible accuracy, feature engineering was applied to deal with outliers. In turn, the interquartile range (IQR) strategy was used to remove outliers above the upper Whiscur value, thus eliminating outliers.

Base models and optimization: to identify the base models, the LR, RF, DT, GNB and KNN algorithms were trained. As the case study seeks to anticipate the presence of prostate cancer, higher priority was given to the sensitivity metric to ensure greater accuracy. The DT and RF models obtained the best scores in sensitivity and ROC score, with 0.81 and 0.67, respectively, so these will be the base models. For optimization, the imbalance that exists in the diagnostic logs was considered, and the SMOTE overbalance technique was applied to mitigate the imbalance. Subsequently, the recursive feature elimination (RFE) technique was implemented for feature selection, then hyperparameter tuning, and finally ensemble techniques were applied.

4. Results and Discussion

During the training stage of the models, DT, RF, AdaBoost and, GB models were used, the data set provided by Kaggle was used. Next, models were developed and refined. Finally, the training results were contrasted under the metrics of recall, sensitivity, precision, and accuracy. The training results are detailed in Table 3.

Table 3. Training Results

Decision Tree	
Train Score	0.9799
Test accuracy	0.9603
F1 score	0.8
Recall	0.9523
Precision	0.6897
Roc auc	0.9567
Random Forest	
Train Score	0.9781
Test accuracy	0.9643
F1 score	0.8085
Recall	0.9048
Precision	0.7308
Roc auc	0.9372
Adaptive Boosting	
Train Score	1

Test accuracy	0.9921
F1 score	0.9524
Recall	0.9524
Precision	0.9425
Roc auc	0.974
Gradient Boost	
Train Score	0.9799
Test accuracy	0.9802
F1 score	0.9524
Recall	0.9524
Precision	0.8333
Roc auc	0.9675

For this study, the training results of the RF, DT, AdaBoost, and GB models were 73.08%, 68.93%, 94.25%, and 83.33% in accuracy, respectively. Regarding sensitivity, the following results were obtained: 90.48%, 95.23%, 95.24%, and 95.25%, respectively.

When examining the results in Table 3, the AdaBoost model has obtained the highest accuracy, with 99.21%, but overfitting has been detected, so it is not an optimal model for cancer prediction. On the other hand, the GB model has obtained 98.02% in accuracy, 95.24% in sensitivity, 88.89% in F1 score, and 83.33% in precision, so it is the best model for cancer prediction. In the second place, we have RF with 96.43% accuracy, 90.48% sensitivity, 80.85% F1 score, and 73.08% accuracy. Finally, the DT model obtained 96.03% accuracy, 95.26% in sensitivity, 80% in F1 score, and 68.97% in precision

ML models have proven to be highly applicable in multiple fields and disciplines, mainly in medicine. Considering that cervical cancer ranks fourth in terms of diagnosis, it is necessary to conduct a study to compare different ML models and identify which one is the most accurate in cervical cancer prediction. The RF, DT, AdaBoost, and GB models were trained, and the findings indicated that the GB algorithm obtained the highest scores, 98.02% in accuracy and 83.33% in precision, these results are similar to those obtained in the study[39], where they concluded that, among the models with better performance, were RF, DT, AdaBoost, and GB, with 100% accuracy, the difference with this research lies in that AdaBoost did not present overfitting in the model. In contrast, the results of [40], positioned the DT and RF models as the most accurate in the prediction of cancer, with 95%. These results surpass those obtained in this study, since these models got an accuracy of 96.03% and 96.43%, respectively, but 68.97% and 73.08% in precision. Similarly, [41] determined that LR and RF are the most accurate models since they achieved 91.4%. Similarly, [42], [43] determined that the RF model is the most accurate, since it got the highest accuracy, 95.68% and 99.8%, in their respective investigations. Regarding the RF and DT algorithms, the latter studies achieved higher results than those of this research, one of the factors that generally influence this result is the quality of the dataset or the optimization methods that were used. The use of ML for cervical cancer prediction can be a valuable weapon to save women's lives, but it is important to emphasize that these algorithms are conditioned to the datasets used for training.

5. Conclusion

Regarding the training results of the RF, DT, AdaBoost, and GB models, we reached the following conclusions.

The GB model obtained the best metrics in precision, sensitivity, and accuracy, making it the indicated algorithm for prostate cancer prediction. Although the AdaBoost algorithm achieved 94.25% accuracy, overfitting was observed, which prevented it from being the best model. On the other hand, according to Fig. 6, poor sexual habits and smoking are factors that influence the probability of cervical cancer. Therefore, it is essential to raise awareness of these factors to prevent many deaths associated with this disease.

The data set presented significant imbalances that could hurt the training of the models; we overcame these drawbacks by applying imputations and sampling techniques. Similarly, the sensitivity metric and ROC score were important factors, since we sought to obtain the highest accuracy in the prediction of the occurrence of prostate cancer, after applying the optimization methods and techniques we were able to increase the sensitivity from 53% to 95.2%.

Finally, we believe that these ML models will be a key factor in prostate cancer prediction, as they will save many lives with an early cancer diagnosis. For future projects, multiple ML algorithms and more datasets should be considered to identify the most efficient algorithm for prostate cancer prediction

This research aims to compare the accuracy of ML models in predicting prostate cancer. The DT, RF, AdaBoost, and GB models are described and analyzed. To determine which model achieves the highest accuracy, DT, RF, LR, GNB, and KNN models are used to identify the base model for algorithm optimization.

Author contributions

Isaac Conde: Conceptualization, Methodology **Saul Beltozar:** data collection, Writing-Original draft preparation **Enrique Diaz:** Investigation, Writing-Reviewing and Editing. **Raul Tejeda:** data collection, Writing-Original draft preparation

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] H. Sung et al., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," CA Cancer J Clin, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[2] World Health Organization, "Cáncer." Accessed: Sep. 24, 2023. [Online]. Available: https://www.who.int/es/news-room/fact-sheets/detail/cancer

[3] Pan American Health Organization, "Día Mundial contra el Cáncer 2023: Por unos cuidados más justos - OPS/OMS | Organización Panamericana de la Salud." Accessed: Sep. 24, 2023. [Online]. Available: https://www.paho.org/es/campanas/dia-mundial-contra-cancer-2023-por-unos-cuidados-mas-justos

[4] F. Yang et al., "Global patterns of cancer transitions: A modelling study," Int J Cancer, Nov. 2023, doi: 10.1002/IJC.34650.

[5] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA Cancer J Clin, vol. 68, no. 6, pp. 394–424, Nov. 2018, doi: 10.3322/caac.21492.

[6] S. Mahjoub and A. Heidenreich, "Oligometastatic prostate cancer: definition and the role of local and systemic therapy: a narrative review," Transl Androl Urol, vol. 10, no. 7, pp. 3167–3175, Jul. 2021, doi: 10.21037/TAU-20-1033.

[7] J. J. Tosoian, M. A. Gorin, A. E. Ross, K. J. Pienta, P. T. Tran, and E. M. Schaeffer, "Oligometastatic prostate cancer: Definitions, clinical outcomes, and treatment considerations," Nat Rev Urol, vol. 14, no. 1, pp. 15–25, Jan. 2017, doi: 10.1038/NRUROL.2016.175.

[8] F. Algaba, "Consideraciones anatomopatológicas a la definición de cáncer de próstata indolente y clínicamente insignificante," Arch Esp Urol, vol. 67, no. 5, pp. 393–399, 2014.

[9] F. Khani et al., "Evolution of structural rearrangements in prostate cancer intracranial metastases," NPJ Precis Oncol, vol. 7, no. 1, Dec. 2023, doi: 10.1038/S41698-023-00435-3.

[10] L. A. Mucci, K. M. Wilson, M. A. Preston, and E. L. Giovannucci, "Is Vasectomy a Cause of Prostate Cancer?," J Natl Cancer Inst, vol. 112, no. 1, pp. 5–6, Jan. 2020, doi: 10.1093/JNCI/DJZ102.

[11] H. J. Chang et al., "A matched case-control study in Taiwan to evaluate potential risk factors for prostate cancer," Sci Rep, vol. 13, no. 1, Dec. 2023, doi: 10.1038/S41598-023-31434-W.

[12] M. Leapman, S. B. Jazayeri, M. Katsigeorgis, A. Hobbs, and D. B. Samadi, "Patient-perceived Causes of Prostate Cancer: Result of an Internet-based Survey," Urology, vol. 99, pp. 69–75, Jan. 2017, doi: 10.1016/J.UROLOGY.2016.09.046.

[13] A. B. Porcaro et al., "Endogenous testosterone density associates with predictors of tumor upgrading and disease progression in the low through favorable intermediate prostate cancer risk categories: analysis of risk factors and clinical implications," African Journal of Urology, vol. 29, no. 1, Jun. 2023, doi: 10.1186/S12301-023-00366-2.

[14] G. S. Dite, E. Spaeth, N. M. Murphy, and R. Allman, "Development and validation of a simple prostate cancer risk prediction model based on age, family history, and polygenic risk," Prostate, vol. 83, no. 10, pp. 962–969, Jul. 2023, doi: 10.1002/PROS.24537.

[15] K. Hemminki, X. Li, A. Försti, and C. Eng, "Are population level familial risks and germline genetics meeting each other?," Hered Cancer Clin Pract, vol. 21, no. 1, Dec. 2023, doi: 10.1186/S13053-023-00247-3.

[16] M. Oderda et al., "Predictors of Prostate Cancer at Fusion Biopsy: The Role of Positive Family History, Hypertension, Diabetes, and Body Mass Index,"

Current Oncology, vol. 30, no. 5, pp. 4957–4965, May 2023, doi: 10.3390/CURRONCOL30050374.

[17] N. Sayegh et al., "Race and Treatment Outcomes in Patients With Metastatic Castration-Sensitive Prostate Cancer: A Secondary Analysis of the SWOG 1216 Phase 3 Trial," JAMA Netw Open, vol. 6, no. 8, p. e2326546, Aug. 2023, doi: 10.1001/JAMANETWORKOPEN.2023.26546.

[18] A. C. Powell, C. T. Lugo, J. T. Pickerell, J. W. Long, B. A. Loy, and A. J. Mirhadi, "An assessment of the association between patient race and prior authorization program determinations in the context of radiation therapy," Healthcare, vol. 11, no. 3, Sep. 2023, doi: 10.1016/J.HJDSI.2023.100704.

[19] S. A. Kaplan, "Benign Prostatic Hyperplasia," Journal of Urology, vol. 210, no. 2, pp. 360–362, Aug. 2023, doi: 10.1097/JU.0000000000003522.

[20] M. J. Arnold, A. Gaillardetz, and J. Ohiokpehai, "Benign Prostatic Hyperplasia: Rapid Evidence Review," Am Fam Physician, vol. 107, no. 6, pp. 613–622, Jun. 2023.

[21] J. Shi et al., "Low-dose antimony exposure promotes prostate cancer proliferation by inhibiting ferroptosis via activation of the Nrf2-SLC7A11-GPX4 pathway," Chemosphere, vol. 339, Oct. 2023, doi: 10.1016/J.CHEMOSPHERE.2023.139716.

[22] O. Bede-Ojimadu et al., "Cadmium exposure and the risk of prostate cancer among Nigerian men: Effect modification by zinc status," Journal of Trace Elements in Medicine and Biology, vol. 78, Jul. 2023, doi: 10.1016/J.JTEMB.2023.127168.

[23] L. Depotte et al., "Association between overweight, obesity, and quality of life of patients receiving an anticancer treatment for prostate cancer: a systematic literature review," Health Qual Life Outcomes, vol. 21, no. 1, Dec. 2023, doi: 10.1186/S12955-023-02093-2.

[24] A. Luciani, C. Falci, F. Petrelli, and G. Colloca, "Prostate Cancer in Older Adults with Frailty," Frailty in Older Adults with Cancer, pp. 357–370, Jan. 2022, doi: 10.1007/978-3-030-89162-6_20.

[25] International Agency for Research on Cancer, "Prostate Source: Globocan 2020 Number of new cases in 2020, both sexes, all ages," 2020, Accessed: Sep. 24, 2023. [Online]. Available: https://gco.iarc.fr/today

[26] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA Cancer J Clin, vol. 68, no. 6, pp. 394–424, Nov. 2018, doi: 10.3322/caac.21492.

[27] F. Hohman, K. Wongsuphasawat, M. B. Kery, and K. Patel, "Understanding and Visualizing Data Iteration in Machine Learning," Conference on Human Factors in Computing Systems - Proceedings, Apr. 2020, doi: 10.1145/3313831.3376177.

[28] X. Borrat, L. A. Celi, and C. Ferrando, "Técnicas Big data para el uso secundario de datos clínicos para la creación de conocimiento medico. La solución MIMIC," Rev Esp Anestesiol Reanim, vol. 66, no. 10, pp. 555–558, Dec. 2019, doi: 10.1016/J.REDAR.2019.07.004.

[29] P. Samuel, Reshmy A. K., S. Rajesh, Kanipriya M., and Karthika R. A., "AI-Based Big Data Algorithms and Machine Learning Techniques for Managing Data in E-Governance," AI, IoT, and Blockchain Breakthroughs in E-Governance, pp. 19–35, May 2023, doi: 10.4018/978-1-6684-7697-0.CH002.

[30] Z. Amiri, A. Heidari, N. J. Navimipour, M. Unal, and A. Mousavi, "Adventures in data analysis: a systematic review of Deep Learning techniques for pattern recognition in cyber-physical-social systems," Multimed Tools Appl, 2023, doi: 10.1007/S11042-023-16382-X.

[31] A. Afzal et al., "Use of modern algorithms for multi-parameter optimization and intelligent modelling of sustainable battery performance," J Energy Storage, vol. 73, Dec. 2023, doi: 10.1016/J.EST.2023.108910.

[32] I. Popchev and D. Orozova, "Algorithms for Machine Learning with Orange System," International journal of online and biomedical engineering, vol. 19, no. 4, pp. 109–123, 2023, doi: 10.3991/IJOE.V19I04.36897.

[33] M. Terra, M. Baklola, S. Ali, and K. El-Bastawisy, "Opportunities, applications, challenges and ethical implications of artificial intelligence in psychiatry: a narrative review," Egypt J Neurol Psychiatr Neurosurg, vol. 59, no. 1, Jun. 2023, doi: 10.1186/S41983-023-00681-Z.

[34] X. Qiao et al., "MRI Radiomics-Based Machine Learning Models for Ki67 Expression and Gleason Grade Group Prediction in Prostate Cancer," Cancers (Basel), vol. 15, no. 18, p. 4536, Sep. 2023, doi: 10.3390/cancers15184536.

[35] X. Deng et al., "Machine learning model for the prediction of prostate cancer in patients with low prostate-specific antigen levels: A multicenter retrospective analysis," Front Oncol, vol. 12, p. 985940, Aug. 2022, doi: 10.3389/fonc.2022.985940.

[36] R. Li, J. Zhu, W. De Zhong, and Z. Jia, "Comprehensive Evaluation of Machine Learning Models and Gene Expression Signatures for Prostate Cancer Prognosis Using Large Population Cohorts," Cancer Res, vol. 82, no. 9, pp. 1832–1843, May 2022, doi: 10.1158/0008-5472.CAN-21-3074.

[37] S. J. Hectors et al., "Magnetic Resonance Imaging Radiomics-Based Machine Learning Prediction of Clinically Significant Prostate Cancer in Equivocal PI-RADS 3 Lesions," Journal of Magnetic Resonance Imaging, vol. 54, no. 5, pp. 1466–1473, Nov. 2021, doi: 10.1002/jmri.27692.

[38] M. M. Hassan et al., "A comparative assessment of machine learning algorithms with the Least Absolute Shrinkage and Selection Operator for breast cancer detection and prediction," Decision Analytics Journal, vol. 7, p. 100245, Jun. 2023, doi: 10.1016/j.dajour.2023.100245.

[39] N. Al Mudawi and A. Alazeb, "A Model for Predicting Cervical Cancer Using Machine Learning Algorithms," Sensors 2022, Vol. 22, Page 4132, vol. 22, no. 11, p. 4132, May 2022, doi: 10.3390/S22114132.

[40] U. K. Lilhore et al., "Hybrid Model for Detection of Cervical Cancer Using Causal Analysis and Machine Learning Techniques," Comput Math Methods Med, vol. 2022, 2022, doi: 10.1155/2022/4688327.

[41] M. S. Al-Batah, M. Alzyoud, R. Alazaidah, M. Toubat, H. Alzoubi, and A. Olaiyat, "EARLY PREDICTION OF CERVICAL CANCER USING MACHINE LEARNING TECHNIQUES," Jordanian Journal of Computers and Information Technology, vol. 8, no. 4, pp. 357–369, Dec. 2022, doi: 10.5455/JJCIT.71-1661691447.

[42] S. K. Suman and N. Hooda, "Predicting risk of Cervical Cancer : A case study of machine learning," Journal of Statistics and Management Systems, vol. 22, no. 4, pp. 689–696, May 2019, doi: 10.1080/09720510.2019.1611227.

[43] R. Alsmariy, G. Healy, and H. Abdelhafez, "Predicting Cervical Cancer using Machine Learning Methods," International Journal of Advanced Computer Science and Applications, vol. 11, no. 7, pp. 173–184, 2020, doi: 10.14569/IJACSA.2020.0110723.