Breast Cancer risk estimation using data mining techniques at a public Mexican hospital

Carlos Eduardo Sánchez Torres
Facultad de Ciencias
Universidad Autónoma de Baja California (UABC)
Ensenada, Baja California
Email: a361075@uabc.edu.mx

***Abstract*—We predict the Breast Cancer risk by applying Random Forest and Gradient Boosting Tree (ensembles learning methods CART, ID3, C4.5) to risk factor datasets and compare their performance to each other.**

I. Introduction

A. Problem

Since breast cancer is the first dead cause in Mexico among women it became a big public health problem –in fact, 2.26 million cases worldwide [17] and [3]. In other words, is a type of cancer with the highest incidence and mortality in women: every day at least 14 women, chiefly between 50 to 69 years, die [5]. Indeed, as we can see in figure 1, breast cancer is an increasing tendency compared to other cancers.

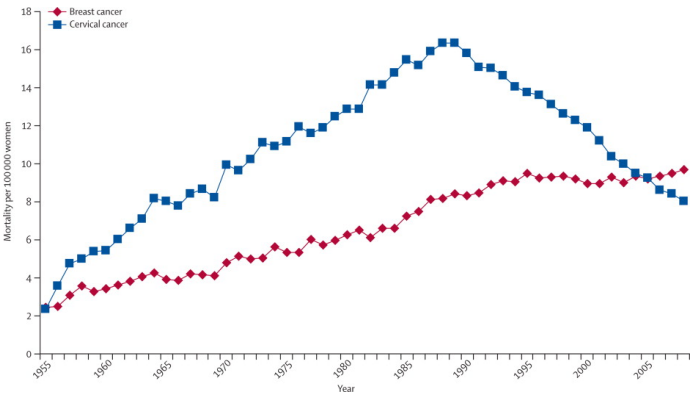


Fig. 1. Breast cancer in Mexico: a challenge to health [3]

Nowadays, doctors carried out analyses using Traditional 2D mammograms from patient requests at public Mexican hospitals. In Ensenada, doctors request private external assistance. Oncologists annotate medical images, and they chiefly say what is the patient's BI-RADS score. "BI-RADS" means Breast Imaging Reporting and Database System, and it's a scoring standard radiologists and oncologists use to describe mammogram results. We'll explain further BI-RADS in section IV.

Our goals are to build data mining models that understand mammograms and predict breast cancer developing risk, continuing [16] and [6] works. Since our model output is a person's future healthy situation, we'll do descriptive and predictive methods, indeed we're going to apply Machine Learning algorithms to datasets. Of course, we don't expect to replace medical doctors but assist them. We know other computer-aided detection systems have been developed in breast cancer detection such as [10] [15], but no one applies them to regional cities and they are not free.

We expect our project can help thousands of women in quickly cancer detection because data mining is faster and cheaper than humans, therefore we're contributing to the decrease in the death rate.

II. Related work

Today, the most common way that machine learning researchers studies mammograms is neuronal network models. [1] developed a deep Convolutional Neural Network (CNN) that classifies traditional 2D mammograms from DDSM into five instances: Normal, Benign Calcification, Benign Mass, Malignant Calcification, Malignant Mass. [12] developed a deep learning project to classify breast cancer into BIRADS Standard from CEDM Images. Since labeling medical images are highly time-consuming, [2] put forward to build a model from transfer learning, and supervised learning on a large labeled dataset and fine-tuned on in-domain medical data.

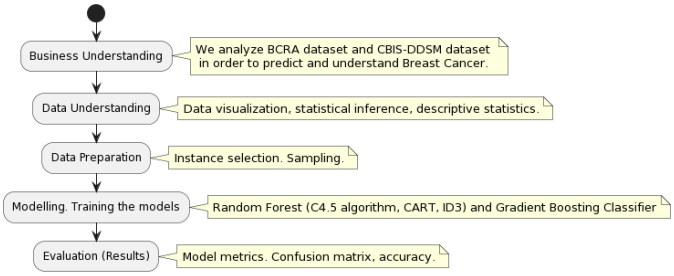
Data miners have used other techniques from other datasets. [16] developed a web system that estimates the suspicion of breast cancer using a gradient-boosting decision tree model from risk factors. [15] developed a system that walks around the major data mining over mini-MIAS database steps: image preprocessing, feature extraction, and the classification process. [10] purposed the BI-RADS feature extraction algorithm for clinical data mining from free text.

III. Methodology

Since we're going to build a Machine Learning model based on data, we have to perform previous exploratory data analysis but not data preprocessing because our datasets

check high standards. We'll use a Jupyter notebook[1] in order to make our project reproducible, easy to change, and self-contained –indeed, since you can test our results on Google Colaboratory, you don't have to download anything. We'll describe specialized tools in the corresponding section.

Our overall workflow compares different algorithms or techniques to get better results, so we'll describe them in great detail in their corresponding section. Our workflow is CRISP-DM process-based [18]. We present you our workflow overview in figure III.



IV. Dataset description

Our model use the well-known statistical model Gail Model, named after Dr. Mitchell Gail, this model use risk factors over specific period of time such as age, race, previous breast biopsies, age at menarche, age at first live birth of a child, and first-degree relatives (mothers, sisters, daughters) with breast cancer, presence of atypical hyperplasia in a biopsy [8], it is described on [19]. We're specially interested on BIRADS breast density, it is our label or class attribute.

However, even tough a woman's risk is high, it doesn't mean that she will develop breast cancer. Hence, women's 45 or older must take an annual mammograms.

The Breast Cancer Risk Assessment (BCRA) is a dataset and algorithms based on Gail model according to National Cancer Institute's Breast Cancer Risk Assessment Tool [19] and [4] that consists 1 522 340 instances and 13 attributes.

On the one hand, the BCRA consists of the categorical values in the period from 2005 to 2017, for example, age groups from 35 to 84 in 4-year intervals. BIRADS breast density code, our class attribute, take values 1, 2, 3, 4, and 9 where increasing values mean more density except for the "9" code, which indicates an unknown measure. [11] explores further details on BIRADS.

[1]We publish our experiments -Jupyter notebook- online on https://gist.github.com/sanchezcarlosjr/bfdbf294e8a89e81c005ac9f8a74a413

On the other hand, model's another input is mammograms which are X-ray pictures. In production, hospitals will provide them by uploading our system. But, in training, we're using CBIS-DDSM [9], a curated mammography dataset for computer-aided detection systems (CADe). CBIS-DDSM consists 2 620 scanned film mammography studies containing normal, benign, and malignant cases. Also, it includes updated mass segmentation, bounding boxes, and pathological diagnosis.

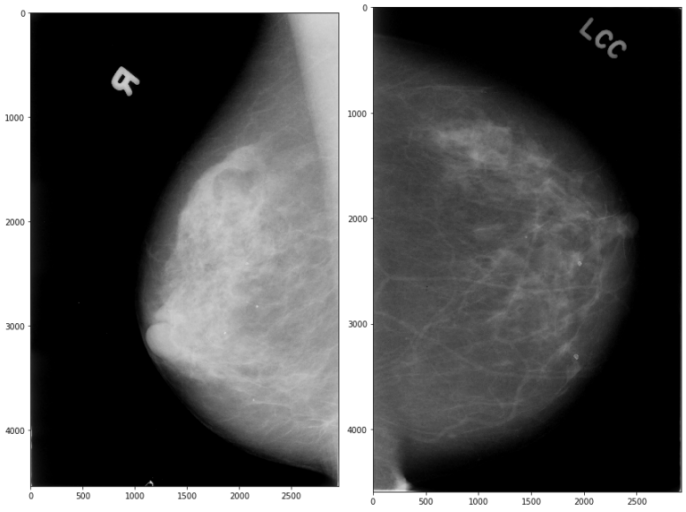


Fig. 2. Images of Craniocaudal - Top View and Mediolateral oblique - Side View from [9]

V. Exploratory data analysis

A. Data visualization

First off, we present you representative charts:

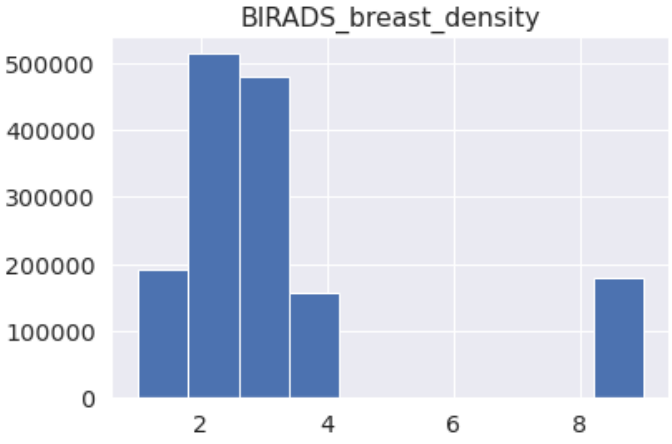


Fig. 3. BIRADS breast density.

BIRADS breast density values are 1, 2, 3, and 9 where 1 means fatty breast, 2 average density, 3 heterogeneously dense, 4 extremely dense, 9 unknown. Histogram 3 has mean and standard deviation 3.21 ± 2.26 density. Indeed, the largest percentage of the population of women with BIRADS breast

density 2 (heterogeneously dense) is 33.76%. The margin of sampling error is ±0.0751% with a 95% level of confidence.

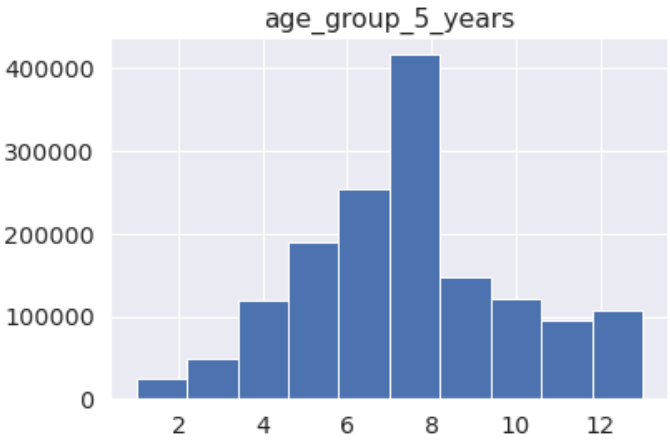


Fig. 4. Age group 5 years.

Age group 5 years values are 1 = 35-39; 2 = 40-44; and so forth, until 13 = 80-84. Histogram 4 has mean and standard deviation 7.299 ± 2.55 group, so the largest percentage of the population of women with age 7 (65-69) is 16.69%. The margin of sampling error is ±0.059% with a 95% level of confidence.

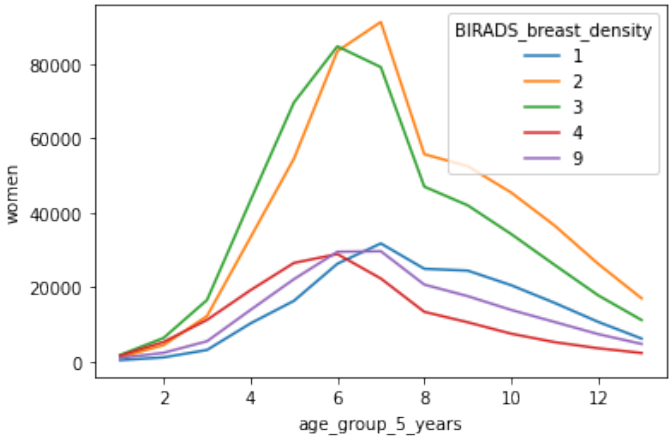


Fig. 5. Age vs BIRADS.

Since these variables check Chi-square Test preconditions, we set the null hypothesis BIRADS and age are not correlated among them; we set the alternate hypothesis as BIRADS and age are correlated between them.

After we applied Chi-Square test, the p-value is 0. Because our level of confidence is 95%, p-value $< \alpha = 0.05$, therefore BIRADS and age are correlated between them.

BMI and BIRADS check the Chi-Square test, we set similar null hypothesis and alternative hypothesis. We apply the Chi-Square test, we get the p-value is 0. Since our level of confidence is 95%, p-value $< \alpha = 0.05$, therefore BMI and BIRADS are correlated with each other.

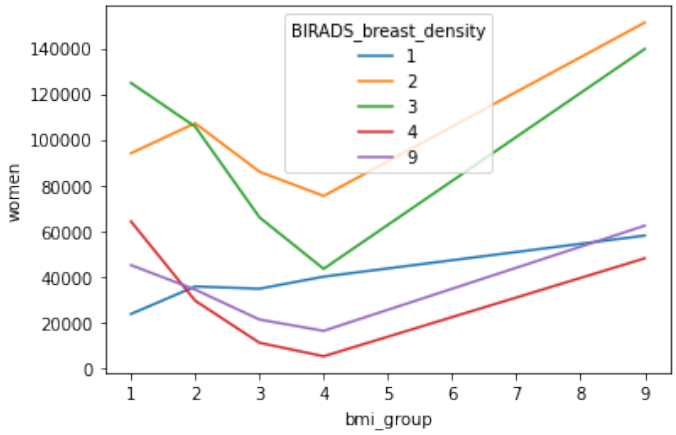


Fig. 6. BMI vs BIRADS.

VI. Prediction Algorithms

We purpose a tool that assists doctors in detecting breast cancer at General Hospital in Ensenada using Random Forest [7] with Gini or Entropy. After we compare Gradient Tree Boosting, other ensemble learning [7]. In both, their inputs are 20% as the testing dataset and 80% as the training dataset with all features except the year, where $BIRADS$ is the label or class. Our project uses Scikit Learn [13] as a Machine Learning library.

Both methods create decision trees, models that predict the value of a class by learning decision rules from the dataset.

Confusion matrices have labeled ”0” means BIRADS 1, ”1” means BIRADS 2, ”2” means BIRADS 3, ”3” means BIRADS 4, and ”4” means unknown.

A. Random Forest

Random Forest is a Supervised Machine Learning Algorithm, or more technically an Ensemble Learning method. Random forests can be regressors or classifiers. We focus on classifiers.

Random forest generates different decision tree classifiers estimating with a criterion its sub-samples quality and improving the general accuracy with averaging. So, two important parameters are the tree algorithm and the criterion.

The Skit Learn case, to generate decision tree classifiers uses an optimized version of the CART algorithm. It supports the criteria of Gini impurity, log loss, and entropy. We experiment with Gini impurity and entropy, but entropy gave us better results. Entropy is a Shannon information gain.

B. Gradient Tree Boosting

Idem that Random Forest, Gradient Tree Boosting is an Ensemble Learning method and we focus on classifiers.

Gradient boosting builds a strong prediction decision tree from the previous weak decision tree, where the current model predicts the error left over by them. It usually outperforms random forest [14].

The Skit Learn Gradient Boosting Classifier generates decision tree classifiers using the AdaBoosting Algorithm. We experiment with a learning rate of 0.1, and the squared loss error function, and 100 estimators. Estimators mean the number of trees in the forest.

VII. Results

Models will validate with the 20% test dataset mentioned previously.

We're going to compare the models and we'll present their metrics.

The feature importance depends on the relative deep in the decision tree with respect to the predictability of the target variable. More above in the tree means more importance.

A. Random forest

A criterion (Gini or Entropy) is an important parameter in a Random forest, we varied it. We present the most important features in figure 9.

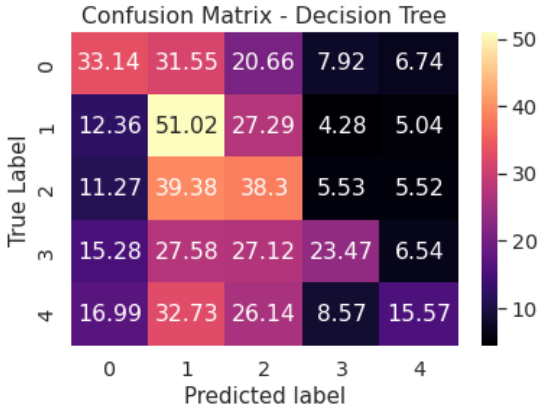


Fig. 7. Random Forest. Criterion Gini. Accuracy 37.38%.

B. Gradient tree boosting

It outperforms the random forest results, but both give low accuracy. They rank similar importance features.

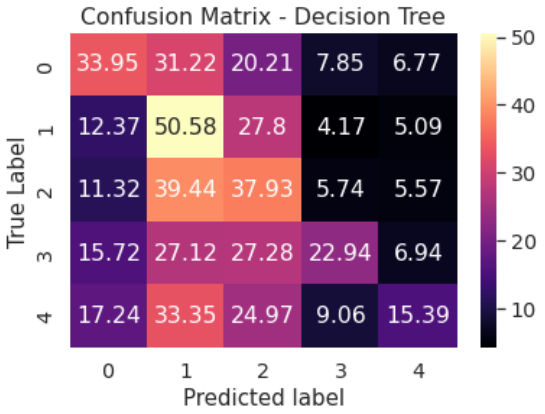


Fig. 8. Random Forest. Criterion Entropy. Accuracy 37.57%

	Feature	Importance
0	age_group_5_years	0.20006562331742875
10	count	0.16322951944714964
4	age_first_birth	0.1415573586689553
3	age_menarche	0.09399111207184864
1	race_eth	0.07263271928824795
6	menopaus	0.06719316432491206
7	bmi_group	0.06461325869532743
8	biophx	0.05842672210074176
2	first_degree_hx	0.05762835867903427
5	current_hrt	0.04198341675611908
9	breast_cancer_history	0.038678746650235046

Fig. 9. Entropy calculated in order to obtain the most discriminative features in Random Forest.

VIII. Conclusions and Limitations

Future work is going to bring out a free computer-aided detection system but with better methods. A big drawback is our analysis is focused on data risk factors rather than film mammographies, but a realistic system has to consider them too. In future work, we'll analyze mammograms with Deep Learning (Neuronal Networks, Tensorflow, Keras) because data risk factors are not enough to predict or understand breast cancer fully.

Since our methods are inductive-based, we can only suggest that age and BMI are the most important factor to have cancer but with low accuracy. Because BCRA is biased and we don't know if American women's conditions are relevant, our arguments are not strong.

References

[1] Data: Abnormality Detection in Mammography using

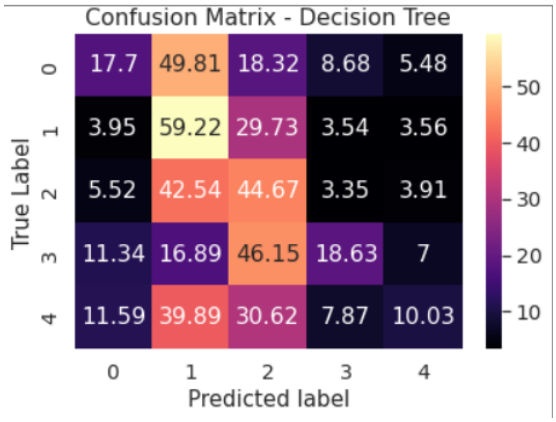


Fig. 10. Gradient tree boosting confusion matrix. Accuracy 39.42%

	Feature	Importance
7	bmi_group	0.2134605162140517
0	age_group_5_years	0.15177684474961067
10	count	0.10425385260104526
3	age_menarche	0.10341563819898514
1	race_eth	0.10189732150847468
4	age_first_birth	0.10096231857538976
9	breast_cancer_history	0.067387016297966
6	menopaus	0.05274271784077392
8	biophx	0.039894578039567724
2	first_degree_hx	0.032883770572919996
5	current_hrt	0.031325425401215176

Fig. 11. Gradient tree boosting feature importance.

Data, November 2022. [Online; accessed 3. Nov. 2022].

[2] Self-Supervised Learning Advances Medical Image Classification, November 2022. [Online; accessed 3. Nov. 2022].

[3] Yanin Chávarri-Guerra, Cynthia Villarreal-Garza, Pedro ER Liedke, Felicia Knaul, Alejandro Mohar, Dianne M Finkelstein, and Paul E Goss. Breast cancer in mexico: a growing challenge to health and the health system. *The Lancet Oncology*, 13(8):e335–e343, 2012.

[4] Data collection and sharing was supported by the National Cancer Institute-funded Breast Cancer Surveillance Consortium (HHSN261201100031C). Risk estimation dataset, October 2020. [Online; accessed 16. Oct. 2022].

[5] Centro Nacional de Equidad de Género y. Salud Reproductiva. Información Estadística Cáncer de Mama, August 2022. [Online; accessed 30. Aug. 2022].

[6] Ramón Santana Fernández, José Manuel Valencia Moreno, and Everardo Gutiérrez López. Risk factors in the appearance of breast cancer, tools, models and current issues. In Xin-She Yang, Simon Sherratt, Nilanjan Dey, and Amit Joshi, editors, *Proceedings of Sixth International Congress on Information and Communication Technology*, pages 869–875, Singapore, 2022. Springer Singapore.

[7] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[8] National Cancer Institute. Breast Cancer Risk Assessment Tool, April 2022. [Online; accessed 7. Oct. 2022].

[9] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4(1):1–9, 2017.

[10] Houssam Nassif, Ryan Woods, Elizabeth Burnside, Mehmet Ayvaci, Jude Shavlik, and David Page. Information Extraction for Clinical Data Mining: A Mammography Case Study. *Proceedings / IEEE International Conference on Data Mining. IEEE International Conference on Data Mining*, page 37, 2009.

[11] S Obenauer, KP Hermann, and E Grabbe. Applications and literature review of the bi-rads classification. *European radiology*, 15(5):1027–1036, 2005.

[12] omar mohamed. Breast-Cancer-Birads-Classification, November 2022. [Online; accessed 3. Nov. 2022].

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[14] S Madeh Piryonesi and Tamer E El-Diraby. Data analytics in asset management: Cost-effective prediction of the pavement condition index. *Journal of Infrastructure Systems*, 26(1):04019036, 2020.

[15] Milos Radovic, Marina Djokovic, Aleksandar Peulic, and Nenad Filipovic. Application of data mining algorithms for mammogram classification. In *13th IEEE International Conference on BioInformatics and BioEngineering*, pages 1–4. IEEE, November 2013.

[16] José Manuel Valencia-Moreno, Everardo Gutiérrez-López, Asley Fernando Cruz González, José Ángel González-Fraga, and José Magaña Magaña. Prototype to estimate breast cancer suspicion in a hospital of the mexican public health system. In *2022 IEEE Mexican International Conference on Computer Science (ENC)*, pages 1–8, 2022.

[17] World Health Organization: Who. Cancer. *World Health Organization: WHO*, February 2022.

[18] Rüdiger Wirth and Jochen Hipp. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical*

applications of knowledge discovery and data mining, volume 1, pages 29–39. Manchester, 2000.

[19] Fanni Zhang. BCRA: Breast Cancer Risk Assessment, October 2022. [Online; accessed 7. Oct. 2022].