Parameter: Journal of Statistics
Available at https://bestjournal.untad.ac.id/index.php/parameter
Vol. 5 No. 2 2025, 76 - 88
DOI: *https://doi.org/10.22487/27765660.2025.v5.i2.17658*

Journal of Statistics
eISSN: 2776-5660

COMPARISON BETWEEN XGBOOST, CATBOOST, RANDOM FOREST, AND LIGHTGBM IN INDONESIAN WOMEN'S BREAST CANCER DATASET

Prajna Pramita Izati[1*], Nuchaila Aniniyah[2], Devi Putri Isnawaty[3]

[1]Universitas Diponegoro
[2,3]Institut Teknologi Sepuluh Nopember

***e-mail**: [1*]*prajnapramitaizati@lecturer.undip.ac.id,* [2]*nuchela24@gmail.com,*[3]*devisnarwaty@its.ac.id*

ABSTRACT

Breast cancer is the most prevalent cancer among women in Indonesia and remains a major public health concern, making the identification of key risk factors essential for early detection. This study applies four machine learning classification algorithms—XGBoost, Random Forest, CatBoost, and LightGBM—to classify breast cancer risk factors using a breast cancer dataset consisting of 400 samples. Data preprocessing was performed prior to analysis, and the dataset was divided into 75% training and 25% testing data using 10-fold cross-validation. Model performance was evaluated using accuracy, precision, recall, F1-score, and area under the curve (AUC). The results show that CatBoost outperforms the other models, achieving the highest AUC value of 0.72. Feature importance analysis indicates that a high-fat diet, menopause status, and working status are the most influential risk factors, while breastfeeding shows a protective effect. These findings demonstrate that CatBoost provides strong predictive performance and effectively identifies key factors associated with breast cancer risk in Indonesia.

Keywords: *Breast Cancer, CatBoost, LightGBM, Random Forest, XGBoost*

Cited: Izati, P. P., Aniniyah, N., & Isnawaty, D. P. (2025). Comparison Between XGBoost, CatBoost, Random Forest, and LightGBM in Indonesian Women's Breast Cancer Dataset. *Parameter: Journal of Statistics, 5(2),* 76-88. https://doi.org/10.22487/27765660.2025.v5.i2.17658

Copyright © 2025 Izati, et al. This is an open-access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

[*]Received: March 2025; Reviewed: Nov 2025; Published: Dec 2025

INTRODUCTION

Breast cancer is a serious threat to women. This is because breast cancer is a type of malignant disease that is very feared after cervical cancer. Breast cancer is an uncontrolled growth of breast cells due to abnormal changes in the genes responsible for regulating cell growth. Normally, old breast cells will die and be replaced by new, more potent cells. This cell regeneration is very useful for maintaining breast function (Putra, 2015). The most common symptom of breast cancer is just a lump. However, in some cases, there are typical symptoms, namely changes in the skin becoming thick with enlarged pores, the nipple turning into the breast, red or blackish brown discharge from the nipple, and blisters on the areola which do not heal easily with ordinary treatment (Tim Edukasi Medis Kanker Payudara, 2017).

Breast cancer is the most prevalent type of cancer in Indonesia and a leading cause of cancer-related mortality, as reported by the Indonesian Ministry of Health. Based on Globocan 2020 data, there were 68,858 new breast cancer cases (16.6%) out of a total 396,914 newly diagnosed cancer cases in the country. Additionally, the number of deaths exceeded 22,000 cases (Rokom, 2022). This represents an important policy priority for the government to be able to reduce or prevent it in overcoming this disease. The important thing to do is diagnose cancer. Cancer diagnosis aims to determine the origin (primary site) of cancer and which cells are involved. Cancer can occur anywhere in the body except for hair, teeth and nails. Cancer diagnosis can be performed through various methods, starting with an initial assessment that includes anamnesis (patient interview) and a physical examination covering the head, eyes, ears, nose, throat, respiratory system, urogenital system, and other body systems. This is followed by laboratory tests to evaluate organ function, along with imaging techniques such as X-rays, MRI, PET scans, CT scans, ultrasound, and endoscopy. Additionally, pathological examinations are conducted to confirm the presence and characteristics of cancer (Kurniasari, S.Gz et al., 2017).

The exact causes of cancer remain unclear, as some types are linked to viral infections. However, according to the World Health Organization (WHO), there are eight established risk factors that contribute to cancer development: obesity and excess body weight, inadequate consumption of fruits and vegetables, lack of physical activity, smoking, alcohol consumption, unsafe sexual practices, air pollution, and aging. Studies conducted in various countries indicate that individuals who are overweight or obese have a 20% higher risk of developing cancer compared to those with a normal body weight (Kurniasari, S.Gz et al., 2017).

According to Ruiz and Hernandez, there is a direct link between diet, lifestyle and risk of developing cancer (Ruiz & Hernandez, 2014). According to Maria, Saina, and Nyorong, several lifestyle risk factors are associated with the incidence of breast cancer, including high fat consumption, obesity, smoking, and stress. Fat intake, particularly saturated fats, is considered a significant risk factor. Consuming foods such as red meat, fried chicken, fast food, full-cream cheese, butter, eggs, and deep-fried foods can increase a woman's likelihood of developing breast cancer. Excessive eating patterns will lead to obesity. Obesity significantly increases the risk of cancer due to the production of estrogen by fat cells. Excess fat cells lead to higher estrogen levels in the body, which can stimulate the growth of cancer cells. Statistical bivariate tests indicate that individuals who frequently consume high-fat foods have a 2.872 times higher risk of developing breast cancer compared to those with lower fat intake. Additionally, obese individuals are 1.942 times more likely to develop breast cancer than those with a normal weight. The probability of developing breast cancer among women who consume high-fat diets and experience stress is estimated to be 65.3% (Maria, Sainal, & Nyorong, 2017). Research conducted by Mohite et al. (2014) examined the link between high fat consumption and obesity as contributing factors to cancer. The findings revealed that dietary risk factors, including excessive intake of visible fat, high salt consumption, a non-vegetarian diet, and being overweight, were significantly associated with an increased incidence of breast cancer among Indian women.

Given the high prevalence of breast cancer among women in Indonesia, it is crucial to classify the factors contributing to its occurrence. Classification is a common technique used to group similar characteristics into specific categories. This study employs four classification methods: eXtreme Gradient Boosting (XGBoost), CatBoost (Categorical Boosting), Random Forest, and Light Gradient Boosting Machine (LightGBM). XGBoost is an advanced gradient tree boosting algorithm designed to efficiently handle large-scale machine learning problems. It was selected for this study due to its enhanced features, which allow for faster computations and reduced risk of overfitting. XGBoost can solve various examples of classification, regression, and ranking. XGBoost is a computational tree collection consisting of various previous trees (CART). The main component behind the prosperity of XGBoost is its adaptability in various situations, this flexibility is due to improvements from past calculations (Yulianti, Soesanto, & Sukmawaty, 2022). In 2017, Ke et al. introduced an algorithm known as the Light Gradient Boosting Machine (LightGBM). This method offers several advantages, including faster and more efficient training, lower memory consumption, improved accuracy, and the ability to effectively process large datasets (Ke, et al., 2017). Random Forest is a machine learning method that is often used in solving supervised problems in machine learning. Random Forest works well with noisy data. One of the uniqueness of CatBoost is the gradient boosting mechanism which is suitable for working with heterogeneous data and can increase stability and good predictive ability. This methodology uses efficient coding and can reduce overfitting (Ray S. , 2020). However, there is still limited research that compares XGBoost, CatBoost, Random Forest, and LightGBM simultaneously for breast cancer classification, especially using data from Indonesia. Most studies focus on a single method, making it difficult to determine which algorithm performs best on heterogeneous medical data. In addition, aspects such as overfitting control, handling categorical variables, and model stability are often overlooked. Therefore, this study addresses this gap by comparing the performance of these four classification methods to identify the most effective approach for breast cancer factor classification.

MATERIALS AND METHODS

This study utilized datasets on reproductive factors, high-fat diet, and body mass index (BMI) as risk factors for breast cancer among Indonesian women. Data were collected from Sardjito Hospital, Yogyakarta, and Dr. M. Djamil General Hospital, Padang. The study involved 200 women diagnosed with breast cancer and 200 women without breast cancer to analyze reproductive factors associated with the disease.

Table 1. Features of Breast Cancer

Demographics	Category	BC(n=200) (%)	Non-BC (n=200) (%)
Age(years)	< 50	47.5	42
	>= 50	52.5	58
Background in education	No degree	1.5	0.5
	Primary school	13	9.5
	Middle school	8	9.5
	High school	40.5	39
	Bachelor's school	34	39.5
	Graduate degree	3	2
working status	Housewife	59	51
	Civil Servant	30	8
	Private Servant	7.5	33
	Enterpreneur	0	3
	Farmer	1	2
	Master's student	0	1
	Retired	6	2
Marital status	Single/widow	15	11.5
	Marriage	18.5	88.5
Age of menarche (years)	7-11	18	16
	12–13	99	91
	> 13	83	93
Age of menopause (years)	>=50	121	79
	< 50	79	121
Age of the first pregnancy (years)	< 20	35	33
	20–29	137	147
	> 30	27	19
	Never been pregnant	1	1
Parity	Nulliparous	1	2
	Primiparous	24	39
	≥ Multiparous	175	159

Demographics	Category	BC(n=200) (%)	Non-BC (n=200) (%)
Breastfeeding	≥ 12 months	198	157
	< 12 months	2	43
High-fat diet	High	192	88
	Normal	8	112
BMI	Normal	87	125
	Normal	38	21
	Obesity	75	54
Ethnicity	Minangnese		
	Javanese		

To accurately classify breast cancer risk factors among Indonesian women, a structured data analysis approach is essential. This study utilizes machine learning classification models to identify patterns and relationships between lifestyle, reproductive factors, and body mass index (BMI) in breast cancer cases. The analysis involves several key steps, including:

1. Initial Data Exploration
   The first step involves exploring the dataset to gain an initial understanding of the breast cancer data. Since the original dataset contains categorical variables in string/object format, these variables are converted into numerical representations for further processing.
2. Handling Missing Data
   A completeness check is performed to identify any missing values. If no missing values are detected, the dataset is considered complete and ready for further analysis.
3. Data Partitioning for Model Training and Evaluation
   The dataset was divided into training and testing subsets using a 75:25 proportion. This ratio was selected because it provides a sufficient amount of data for model training and parameter optimization while preserving an adequate portion of unseen data for an unbiased performance evaluation. The data splitting procedure was carried out using a stratified sampling approach to ensure that the class distribution in both the training and testing sets remained representative of the original dataset. This procedure is described explicitly to guarantee reproducibility and to avoid potential bias in model evaluation.
4. Implementation of Classification Algorithms
   Four machine learning classification models—CatBoost, Random Forest, XGBoost, and LightGBM—are applied to the dataset. Model training and evaluation use repeated holdout validation and 10-fold cross-validation (CV). The 10-fold CV is chosen because it is a standard and widely used setting that provides stable and reliable performance estimates without requiring excessive computation.
   a. XGBoots
   The eXtreme Gradient Boosting (XGBoost) method is an enhanced version of the gradient boosting technique, introduced by Dr. Tianqi Chen from the University of Washington in 2014. XGBoost is built upon the Classification and Regression Trees (CART) algorithm, commonly referred to as Decision Trees (Chen & Guestrin, 2016). XGBoost is a highly scalable, flexible, and versatile machine learning tool designed to efficiently utilize computational resources and address the limitations of previous gradient boosting methods. The key distinction between XGBoost and other gradient boosting techniques lies in its regularization mechanism, which helps mitigate overfitting. This enhancement makes XGBoost faster and more robust during model optimization. The regularization is achieved by incorporating an additional term into the loss function, improving the model's generalization ability, as (Daoud, 2019):

$$L(f) = \sum_{i=1}^{n} L(\hat{y}_i, y_i) + \sum_{m=1}^{M} \Omega(\delta_m) \quad (1)$$

with

$$\Omega(\delta) = a|\delta| + 0.5\beta\|w\|^2 \tag{2}$$

where $|\delta|$ represents the number of branches, w denotes the value of each leaf and Ω refers to the regularization function.

b. CatBoots

CatBoost (Categorical Boosting) is a recently open-sourced machine learning algorithm developed by Yandex. It seamlessly integrates with deep learning frameworks such as Google's TensorFlow and Apple's Core ML, making it adaptable for various applications. Designed to handle diverse data types, CatBoost effectively addresses a wide range of business challenges while delivering state-of-the-art accuracy (Ray S. , 2020). A key feature of CatBoost is its specialized handling of categorical variables, utilizing techniques such as permutation-based encoding, One_Hot_Max_Size (OHMS), and target-based statistics to enhance model performance.

c. Random Forest

The Random Forest algorithm enhances prediction accuracy and mitigates overfitting by leveraging averaging techniques. When bootstrap sampling is enabled (bootstrap=True by default), the sub-sample size is regulated using the max_samples parameter; otherwise, the entire dataset is utilized for constructing each decision tree (Pedregosa, Varoquaux, Gramfort, Michel, & Thirion, 2011). In data classification, Random Forest applies the Gini Index formula to determine how nodes in a decision tree split. This formula evaluates class probabilities to compute the Gini value for each branch at a node, helping to identify the most probable branching path.

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2 \tag{3}$$

Where p_i denotes the relative frequency of each class within the dataset, and c represents the total number of classes. In addition to the Gini Index, entropy is also commonly used to determine how nodes split within a decision tree. The entropy calculation follows the formula presented in the equation below.

$$Entropy = \sum_{i=1}^{c} -p_i * log2\ (p_i) \tag{4}$$

Entropy utilizes the probability of outcomes to determine the optimal way for nodes to branch in a decision tree. Unlike the Gini Index, entropy involves more complex mathematical computations due to the use of a logarithmic function in its calculation (Pedregosa, Varoquaux, Gramfort, Michel, & Thirion, 2011).

d. LightGBM

LightGBM is a fast, distributed, and high-performance gradient boosting framework based on decision tree algorithms, designed for tasks such as ranking, classification, and various other machine learning applications. Fundamentally, LightGBM is an ensemble method that aggregates predictions from multiple decision trees by summing their outputs to produce a final, well-generalized prediction. A key characteristic of LightGBM is its additive training process, where multiple trees are sequentially trained, with each new tree learning to predict the residual errors of the previous model. Suppose a LightGBM model consists of a tree (T) and is applied to a dataset with n examples—the additive training process can be described as follows (Chen, et al., 2019).

$$\begin{gathered} \hat{y}_t^{(t)} = 0 \\ \hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\ \hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\ \vdots \\ \hat{y}_i^{(t)} = \sum_{i=1}^{n} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \end{gathered} \tag{5}$$

where

$\hat{y}_i^{(t)}$ = prediction from the *i*-th example in the *t*-th iteration

f_t = the learned function for the *t*-th decision tree.

The equation above illustrates that, in each iteration, the current model is maintained as $\hat{y}_t$, while a new function *f* (representing the learned residual) is added to improve the model's predictions. The function f_s from all iterations is optimized by minimizing the following objective function (Chen, et al., 2019).

$$L^{(t)} = \sum_{i=1}^{n} l(\,y_i, \hat{y}_i^{(t)}) + \sum_{m=1}^{M} \Omega(f_t) \tag{6}$$

The first term represents the loss function, which quantifies the difference between the predicted value $\hat{y}_i^{(t)}$ and the actual target y_i. The second term is the regularization component, which penalizes model complexity to prevent overfitting.

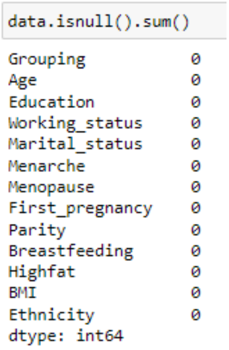
5. Performance Evaluation of Classification Models
   The classification models are evaluated based on their predictive performance using both holdout validation and k-fold cross-validation metrics. Performance indicators such as accuracy, precision, recall, and area under the curve (AUC) are considered in the assessment.
6. Selection of the Optimal Model
   The classification models are compared based on their AUC values, with the model achieving the highest AUC being selected as the best-performing model for breast cancer classification.
7. Interpretation of Key Factors Associated with Breast Cancer
   Finally, the relationships between significant predictive variables and breast cancer status are analyzed. Feature importance scores from the best-performing model are examined to determine the most influential risk factors, providing insights into their impact on breast cancer development.

RESULTS AND DISCUSSION

To compare different gradient boosting methods, the "breast cancer" dataset was utilized and tested by implementing XGBoost, Random Forest, LightGBM, and CatBoost. Prior to classification, the data underwent preprocessing to ensure quality and consistency. Additionally, the characteristics of each variable in the dataset were analyzed to gain deeper insights before applying classification models.

Preprocessing Data Breast Cancer

In data preprocessing, the first step taken is data cleaning. Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. Before visualizing the data, the data cleaning process is first carried out to find out whether there is a missing value in the "Breast Cancer Dataset" or not.

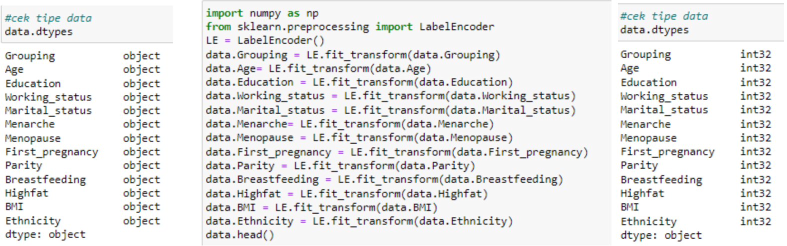


```
data.isnull().sum()

Grouping          0
Age               0
Education         0
Working_status    0
Marital_status    0
Menarche          0
Menopause         0
First_pregnancy   0
Parity            0
Breastfeeding     0
Highfat           0
BMI               0
Ethnicity         0
dtype: int64
```

Figure 1. Checking Missing Value in Dataset

Based on Figure 1. It can be seen that there is no missing value in all features in the dataset. After cleansing the missing value, then checking the data type on each feature. Based on Figure 2, the data types on all variables are still in the form of *objects*, so we need to convert them into *integers* so that modeling can be done.



```
#cek tipe data
data.dtypes

Grouping           object
Age                object
Education          object
Working_status     object
Marital_status     object
Menarche           object
Menopause          object
First_pregnancy    object
Parity             object
Breastfeeding      object
Highfat            object
BMI                object
Ethnicity          object
dtype: object
```

(a)

```
import numpy as np
from sklearn.preprocessing import LabelEncoder
LE = LabelEncoder()
data.Grouping = LE.fit_transform(data.Grouping)
data.Age= LE.fit_transform(data.Age)
data.Education = LE.fit_transform(data.Education)
data.Working_status = LE.fit_transform(data.Working_status)
data.Marital_status = LE.fit_transform(data.Marital_status)
data.Menarche= LE.fit_transform(data.Menarche)
data.Menopause = LE.fit_transform(data.Menopause)
data.First_pregnancy = LE.fit_transform(data.First_pregnancy)
data.Parity = LE.fit_transform(data.Parity)
data.Breastfeeding = LE.fit_transform(data.Breastfeeding)
data.Highfat = LE.fit_transform(data.Highfat)
data.BMI = LE.fit_transform(data.BMI)
data.Ethnicity = LE.fit_transform(data.Ethnicity)
data.head()
```

(b)

```
#cek tipe data
data.dtypes

Grouping           int32
Age                int32
Education          int32
Working_status     int32
Marital_status     int32
Menarche           int32
Menopause          int32
First_pregnancy    int32
Parity             int32
Breastfeeding      int32
Highfat            int32
BMI                int32
Ethnicity          int32
dtype: object
```

(c)

Figure 2. (a) Type of Dataset (b) Transformation of fatures in Dataset using LabelEncoder (c) Type of Data After Transformation Data

In the obtained dataset, there is a feature that has a string data type. Encoder labels are done to convert features that have the form of string data into integer data. For example, the menopause feature has 2 categories of data, namely < 50 years and > 50 years. The category is transformed into values 0 and 1. In addition to the value in the selector/class feature also transformed into 0 and 1.

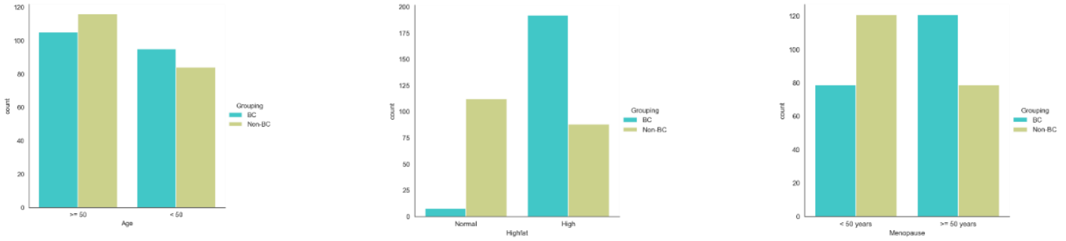
Information About Breast Cancer Dataset

Breast Cancer data consists of 13 features. The variables are of type string or categorical so that to see the characteristics of the data using the mode value.

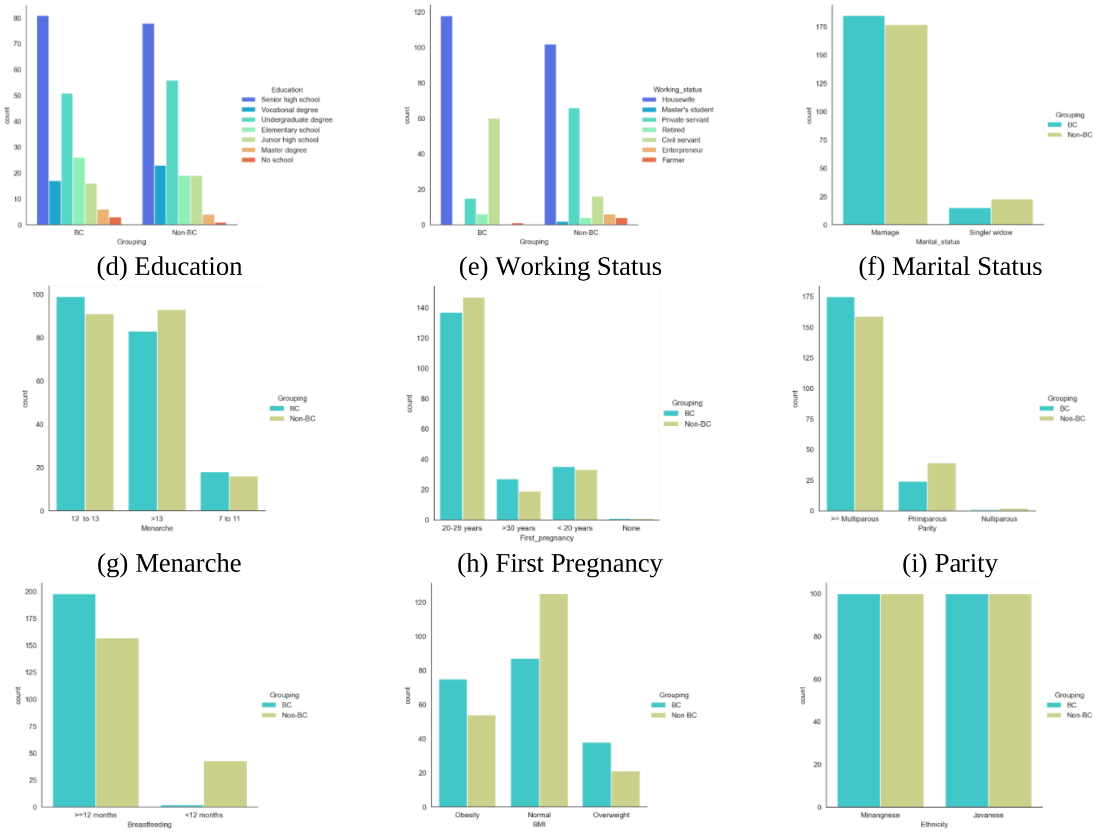
Table 2. The Mode value of each features

Features	Count	Unique	Top	Freq
Grouping	400	2	BC	200
Age	400	2	>= 50	221
Education	400	7	Senior high school	159
Working_status	400	7	Housewife	220
Marital_status	400	2	Marriage	362
Menarche	400	3	12 to 13	190
Menopause	400	2	< 50 years	200
First_pregnancy	400	4	20-29 years	284
Parity	400	3	>= Multiparous	334
Breastfeeding	400	2	>=12 months	355
Highfat	400	2	High	280
BMI	400	3	Normal	212
Ethnicity	400	2	Minangnese	200

From Table 2 shows the value of the mode and the unique number of each feature. The Grouping feature has 2 categories with BC as the most category, which there are 200 respondents suffering from breast cancer. In addition, 221 respondents had an age of more than 50 years. Most respondents were housewives and were married.



(a) Age (b) Highfat (c) Menopause



(d) Education (e) Working Status (f) Marital Status

(g) Menarche (h) First Pregnancy (i) Parity

(j) Breastfeeding (k) BMI (l) Ethnicity

Figure 3. Bar Chart Between Grouping with Each Features

Women over 50 face a higher risk of breast cancer than younger women, with aging and high fat intake being significant risk factors. Menopause after 50 further increases the likelihood, while premenopausal women have a greater risk compared to postmenopausal women of the same age. Education does not significantly influence breast cancer risk, but certain occupations, such as civil servant roles, are associated with a higher risk compared to private sector jobs. Marital status may also play a role, as some studies suggest unmarried women have higher breast cancer rates than married women. Early menarche (≤13 years) and multiple pregnancies are linked to an increased risk, whereas an early first full-term pregnancy (before 30) reduces the likelihood. Breastfeeding provides some protective effects, with the risk decreasing for every 12 months of breastfeeding but potentially increasing after extended periods. Excess body fat raises the risk even in individuals with a normal BMI. While ethnicity generally does not have a significant impact, younger Black and non-Hispanic Black women show higher breast cancer rates compared to their white and non-Hispanic white counterparts.

Classification of Breast Cancer Dataset

Testing the classification model in this experiment will use four classification algorithms are Random Forest, CatBoost, XGboost and LightGBM. From all the models built, the best model will be searched based on the execution results. To determine the best model, the measurement of the classification algorithm model uses the confusion matrix measurement metric.

This study uses two scenarios that will be tried are the first scenario is a scenario that discusses the results obtained by splitting training data and testing data using Training-Testing Repeated Holdout and the second scenario is the results obtained by splitting data using K-Fold Cross Validation.

Data processing with holdout validation techniques where 25% of data is used as testing and the remaining 75% of data as training is repeated to produce the best accuracy. Meanwhile, K-Fold Cross Validation testing uses 10-fold cross validation. Table 3 and Table 4 provide an overview of the results obtained from scenario 1 and scenario 2.

Table 3. The Result of Accuracy Scenario 1

Methods	Accuracy	Precision	Recalls	F1 Score	Runtime Training	Runtime Prediction
Catboost	0,64	0,760417	0,64	0,454545	0,131240	0,001878
XGBoost	0,63	0,642857	0,63	0,564706	0,109271	0,004157
Random Forest	0,62	0,637868	0,62	0,536585	0,296054	0,034506
LightGBM	0,61	0,655280	0,61	0,465753	0,071698	0,005446

Table 4. The Result of Accuracy Scenario 2

Methods	Precision	Recalls	F1 Score	Accuracy	Standar Deviasi Score
Catboost	0,79	0,79	0,79	0,84	0,05333
Random Forest	0,7725	0,7725	0,7725001	0,78667	0,081921
Lightgbm	0,78	0,78	0,78	0,78667	0,080554
XGBoost	0,7575	0,7575	0,7575	0,77667	0,071718

Based on the results presented in Table 3 and Table 4, scenario 2 with a cross-validation (CV) value of 10 demonstrates the best performance. The AUC values for each model are illustrated in Figure 4. The final output data, consisting of 12 features, shows that CatBoost achieves an accuracy of 84%, with a precision, recall, and F1-score of 0.79, and an AUC value of 0.72. As indicated in Table 3, LightGBM exhibits the fastest training time while maintaining high accuracy, whereas Random Forest has the slowest training time. CatBoost, on the other hand, achieves the highest AUC score and the quickest prediction time. However, these findings cannot be universally applied to all datasets. For instance, in datasets with a higher number of categorical features, CatBoost is expected to outperform the other models. Moreover, implementation time appears to be relatively independent and shows a low correlation with feature types.

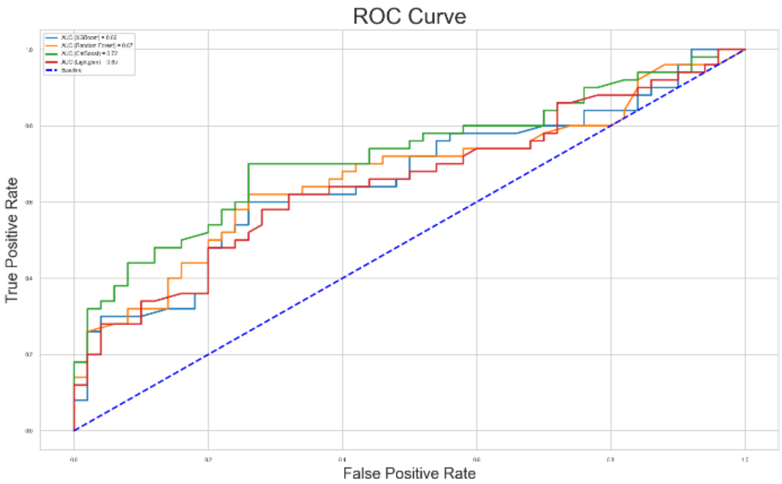
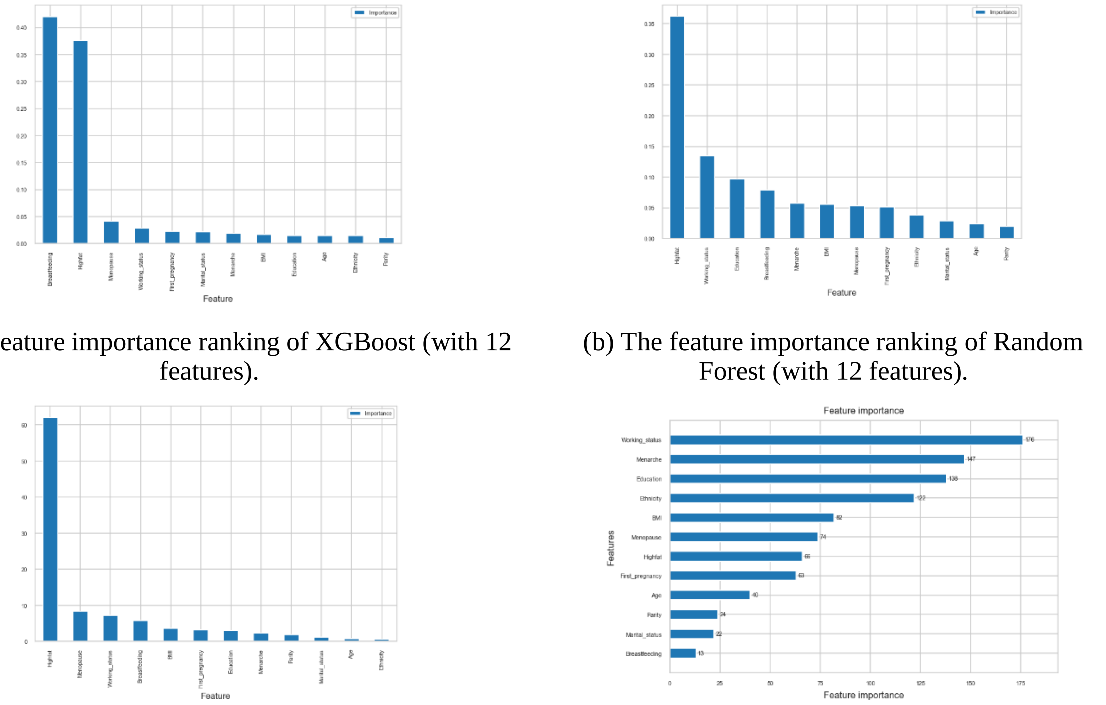


Figure 4. The Result of AUC Value of Each Methods

The CatBoost model has an AUC value of 0.72, the Random Forest model has an AUC value of 0.67, the XGBoost model has an AUC value is 0.66 and the LightGBM model has an AUC value of 0.64. From the results of the model test, the ROC AUC curve can be visualized to present the model performance which is calculated based on the true positive rate and false positive rate data which can be seen in Figure 4.

Next, we conduct a feature importance analysis. Figure 5 (a)-(d) illustrates the ranking of features based on XGBoost, Random Forest, LightGBM, and CatBoost, respectively. As depicted in Figure 5(c), the three most influential features in CatBoost are Highfat, Menopause, and Working Status. Meanwhile, features such as First Pregnancy, Education, Marital Status, Parity, and Menarche hold lower importance but still contribute to the model's performance. Given that some features have minimal impact on breast cancer prediction, we select the top 10 features for further analysis using CatBoost, as it has demonstrated the best performance for this dataset.



(a) The feature importance ranking of XGBoost (with 12 features).

(b) The feature importance ranking of Random Forest (with 12 features).

(c) The feature importance ranking of CatBoots (with 12 features).

(d) The feature importance ranking of LightGBM (with 12 features).

Figure 5. The Result of Features Importances of Each Methods

After applying CatBoost with the top 10 features for breast cancer prediction, we conduct another feature importance analysis based on the trained model. The ranking of feature importance is illustrated in Figure 6. In general, compared to the CatBoost model with 12 features, the ranking of most features remains relatively unchanged, except for Working Status and First Pregnancy, which show slight shifts in importance. As observed in Figure 6, the performance of CatBoost with 10 features is comparable to that with 12 features, showing a slight increase in accuracy mean while experiencing a slight reduction in standard deviation. This suggests that the two removed features have minimal impact on the model and do not significantly contribute to breast cancer prediction.

Table 5. Performance Comparison between methods use 12 features and 10 features

Features	Methods	Mean Score
12	Catboost	0,84
	XGBoost	0,78667
	Random Forest	0,78667
	LightGBM	0,77667
10	Catboost	0,84333
	XGBoost	0.79
	Random Forest	0.79
	LightGBM	0,7933

In this section, we delve deeper into the feature importance analysis. Feature importance measures the contribution of each variable in enhancing the model's predictive capability. It provides an intuitive understanding of which features significantly impact the final model; however, it does not determine the nature of the relationship between a feature and the prediction outcome. As shown in Figure 6, the most influential factors in breast cancer prediction are Highfat, Working Status, and Menopause. However, while these features play a crucial role in the model, Figure 6 does not reveal whether their impact is positive, negative, or follows a more complex pattern in relation to breast cancer prediction.

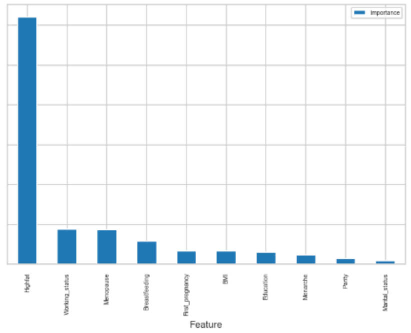
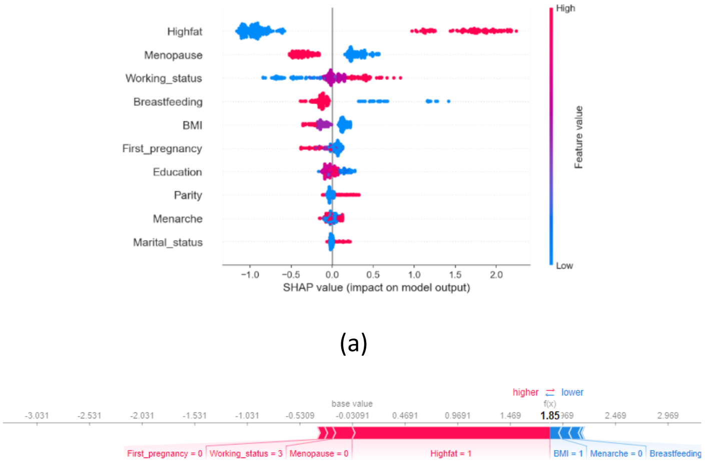


Figure 6. The Result of Features Importances ranking of CatBoost (with 10 features)



(a)

(b)

Figure 7. Analysis of Features use SHAP

SHAP values provide insight into the contribution of each factor in a model's prediction. As an interpretable tool for understanding machine learning model outputs, SHAP not only highlights the influence of features in individual data samples but also distinguishes between their positive and negative effects. In SHAP visualizations, each group of points is color-coded according to feature values, where higher values appear redder. Features are ranked based on their mean SHAP values. Figure 7(b) illustrates the SHAP values for a single data sample, where the base value (-0.03091) represents the mean of the target fitting values in the training set. The visualization also demonstrates how each feature either increases or decreases the predicted value, ultimately resulting in a final prediction of 1.85. Blue indicates a negative contribution, whereas red signifies a positive contribution. A positive SHAP contribution means that a feature increases the predicted value and pushes the model toward the positive class, while a negative contribution means that a feature decreases the predicted value and pushes the model toward the negative class. The final prediction is obtained by combining all positive and negative contributions from the features starting from the base value.

In Figure 7(a), each row represents a feature, with the SHAP value displayed along the horizontal axis, while each point corresponds to a data sample. The color gradient indicates feature magnitude, where red represents higher values and blue represents lower values. From Figure 7(a), it is evident that Highfat is a highly significant feature, showing a strong positive correlation with breast cancer. Other features, such as BMI, first pregnancy, parity, and marital status, also exhibit a positive association, meaning that higher values of these features increase the likelihood of breast cancer prediction. Conversely, breastfeeding is negatively correlated with breast cancer, where lower values contribute to better model predictions. Features like menopause, education, and working status exhibit both positive and negative correlations, suggesting a more complex relationship with breast cancer risk. If a feature shows both positive and negative SHAP values, it means that the feature does not have a uniform effect on the model's prediction. Depending on its value in a particular data sample, the feature can either increase or decrease the predicted risk of breast cancer. This indicates a complex or non-linear relationship, where the effect of the feature varies across individuals rather than influencing all predictions in the same direction.

CONCLUSION

This study applies XGBoost, Random Forest, CatBoost, and LightGBM to classify breast cancer data and evaluates model performance using accuracy, precision, recall, F1-score, and AUC. The Breast Cancer Dataset was preprocessed to ensure data quality, and the data were split into 75% for training and 25% for testing using 10-fold cross-validation. The results show that CatBoost achieves the best performance, with an AUC value of 0.72. The analysis indicates no significant performance difference between models using 12 features and those using 10 features when applying the CatBoost algorithm. Furthermore, the SHAP analysis reveals that a high-fat diet is the most influential risk factor, followed by menopause and working status, while breastfeeding shows a protective effect. These findings confirm that CatBoost not only provides the best predictive performance but also effectively identifies key factors associated with breast cancer risk.

REFERENCES

Bekkar, M., Djemaa, K. D., & Alitouche, A. D. (2013). Evaluation Measures for Models Assessment over Imbalanced Data Sets. *Journal of Information Engineering and Applications*.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System.

Chen, T., Xu, J., Ying, H., Chen, X., Feng, R., Fang , X., Wu, A. J. (2019). Prediction of Extubation Failure for Intensive Care Unit Patienst.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2014). Random Forests.

Daoud, E. A. (2019). Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset. *13*(1).

DQLAB. (2022, September 21). *Studi Kasus Random Forest Machine Learning untuk Pemula Data*. Retrieved Desember 20, 2022, from dqlab.id: https://dqlab.id/studi-kasus-random-forest-machine-learning-untuk-pemula-data

Gokgoz, E., & Subasi, A. (2015). Comparison of Decision Tree Algorithms for EMG Signal Classification using DWT. *18*(138-144).

Han, J., Kamber , M., & Pei, J. (2012). *Data Mining: Concepts and Techniques.* USA: Elsevier Inc.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., . . . Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree.

Kurniasari, S.Gz., MPH, F. N., Harti, S.Gz, MsiMed, L. B., Ariestiningsih, S.Gz., M.P, A. D., Wardhani, SpPD, d. O., & Nugroho, SpA (K), d. (2017). *Buku Ajar Gizi dan Kanker.* Malang: UB Press.

Maria, I. L., Sainal, A. A., & Nyorong, M. (2017). RISIKO GAYA HIDUP TERHADAP KEJADIAN KANKER PAYUDARA PADA WANITA. *13*(2).

Mohite, V. R., Pratinidhi, A. K., & Mohite, R. V. (2014). Dietary factors and breast cancer: A case control study from rural India. *6*(1).

Pedregosa, F., Varoquaux, Gramfort, A., Michel, V., & Thirion, B. (2011). Scikit-Learn: Machine Learning in Python. *12*.

Putra, S. R. (2015). *Buku Lengkap Kanker Payudara.* Yogyakarta: Laksana.

Ray, S. (2020, Juni 07). *Analytics Vidhya*. Retrieved Desember 20, 2022, from CatBoost: A machine learning library to handle categorical (CAT) data automatically: https://www.analyticsvidhya.com/blog/2017/08/catboost-automated-categorical-data/

Ray, S. (2020`, Juni 7). *CatBoost: A machine learning library to handle categorical (CAT) data automatically*. Retrieved Desember 19, 2022, from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2017/08/catboost-automated-categorical-data/

Rokom. (2022, Februari 09). *Kanker Payudara Paling Banyak di Indonesia, Kemenkes Targetkan Pemerataan Layanan Kesehatan*. Retrieved Desember 18, 2022, from Sehat Negeriku, Kementerian Kesehatan: https://sehatnegeriku.kemkes.go.id/baca/umum/20220202/1639254/kanker-payudaya-paling-banyak-di-indonesia-kemenkes-targetkan-pemerataan-layanan-kesehatan/

Ruiz, R. B., & Hernandez, P. S. (2014). Diet and Cancer : Risk Factors and epidemiological evidence. (Maturitas 77).

Saha, S. (2022, November 14). *XGBoost vs LightGBM: How Are They Different*. Retrieved Desember 20, 2022, from Neptune.ai.

Tim Edukasi Medis Kanker Payudara. (2017). *Cerdas menghadapi Kanker Payudara.* Depok: Sinergi Publishing.

Yulianti, S. E., Soesanto, O., & Sukmawaty, Y. (2022). Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit. *4*(1).

Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel. *58*.