PREDICTION OF BRONCHITIS DISEASE INDICATIONS USING THE CATBOOST ALGORITHM

Lukman Hakim[1*]

[1]*Universitas Pembangunan Nasional “Veteran” Jawa Timur (Indonesia)*

*) email : 21081010118@student.upnjatim.ac.id

Abstract

Bronchitis is one of the respiratory diseases classified as Acute Respiratory Infection (ARI), characterized by prolonged cough, shortness of breath, and fever. Accurate prediction of bronchitis indications can assist in early diagnosis and improve the efficiency of healthcare services. This study applies the CatBoost algorithm to predict bronchitis indications based on patient symptom data obtained from an apothecary dataset. The research stages include data collection, data cleaning, labeling, feature engineering, data splitting, hyperparameter tuning using GridSearchCV, model training, and model evaluation. The evaluation was carried out using Accuracy, Precision, Recall, and F1-Score metrics. The results show that the 60:40 data split scenario produced the best performance with an accuracy of 81.81%, precision of 75.23%, recall of 79.66%, and an F1-Score of 77.43%. These findings indicate that the CatBoost algorithm can classify bronchitis indications with good and stable performance.

Keywords: *Machine Learning; CatBoost; Bronchitis Disease Prediction; Classification*.

1. INTRODUCTION

Acute Respiratory Tract Infection (ISPA) is one of the major global health problems that has a significant impact on worldwide morbidity and mortality rates. This disease encompasses a wide range of infectious conditions affecting both the upper and lower respiratory tracts, including the nose, trachea, and lungs. According to reports from the World Health Organization (WHO), ISPA is responsible for approximately 4 million deaths annually worldwide, with a high incidence particularly among children, the elderly, and individuals with weakened immune systems.

In developing countries such as Indonesia, ISPA remains a major burden on the public health system due to its high transmission rate, limited diagnostic facilities, and low public awareness regarding early detection of respiratory diseases. Data from the Ministry of Health of the Republic of Indonesia indicate that ISPA is among the top ten causes of visits to primary healthcare facilities. Approximately 40–60% of visits to community health centers (puskesmas) and 15–30% of outpatient and inpatient hospital visits are associated with ISPA cases. One of the most common forms of ISPA is bronchitis, which is defined as inflammation of the respiratory tract—particularly the trachea and bronchi—caused by viral or bacterial infections. Acute bronchitis is generally characterized by a productive cough lasting more than two weeks, accompanied by shortness of breath, fever, and chest discomfort. Environmental factors such as air pollution, cigarette smoke, and dust exposure further exacerbate this condition, especially in urban areas with high levels of air pollution.

In the era of digitalization and data-driven healthcare technology development, machine learning (ML) methods have emerged as promising approaches for disease analysis and prediction. These techniques enable computers to learn from historical data and identify patterns that are difficult to detect manually by humans. By utilizing patient symptom data, medical history, and medication consumption behavior, machine learning algorithms can construct predictive models that support early diagnosis and evidence-based medical decision-making.

One algorithm that has demonstrated high performance in healthcare data classification is CatBoost (Categorical Boosting), developed by Yandex Research. CatBoost is a variant of the gradient boosting algorithm specifically designed to efficiently handle categorical features without requiring complex manual transformations such as label encoding or one-hot encoding. The main advantage of CatBoost lies in its ability to prevent target leakage through the Ordered Target Statistics (OTS) mechanism, as well as its use of symmetric decision trees, which enhance training stability and prediction accuracy.

Compared to other algorithms such as XGBoost and LightGBM, CatBoost has shown more consistent performance, particularly on datasets with limited size but a high proportion of categorical variables. Furthermore, CatBoost offers high computational efficiency and strong generalization performance on test data, making it an effective algorithm for the development of data-driven medical prediction systems.

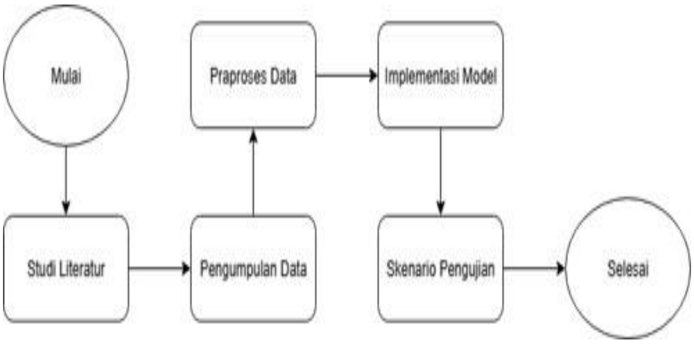
Based on this background, this study aims to apply the CatBoost algorithm to predict indications of bronchitis as a form of ISPA using patient symptom data. Through preprocessing, feature engineering, and hyperparameter optimization, the resulting model is expected to provide accurate predictions and support early screening of bronchitis in a data-driven manner.

2. METHODOLOGY

This study employs a quantitative research method with a supervised learning approach, aiming to develop a classification model to predict indications of bronchitis based on patient symptom data. This approach is selected because it enables systematic and measurable analysis of the relationship between input variables (symptom features) and the output variable (disease indication labels).

The research process is conducted through several main stages, as illustrated in **Figure 1**.

Figure 1. Research Flowchart



The research stages begin with data collection, followed by data preprocessing (data cleaning, feature engineering, and labeling). Subsequently, the dataset is divided into several ratios (60:40, 70:30, and 80:20) to determine the optimal combination of training and testing data. The best ratio is selected based on the model performance evaluation results and is then used in the final training stage of the CatBoost model optimized using GridSearchCV. The final stage is model evaluation, which is conducted using various performance metrics to assess the model's ability to classify bronchitis indications.

2.1. Data Collection

The data used in this study are secondary data obtained from an pharmacy information system, containing records of patient symptoms and prescribed medications related to respiratory tract diseases. The dataset includes several important attributes such as cough+, fever, shortness of breath, sore throat, patient age, and gender. The data were collected during the period of January–March 2025, totaling 2,000 records. The dataset was then filtered to ensure that only data relevant to bronchitis cases were included in the analysis.

2.2. Data Preprocessing

The preprocessing stage is conducted to prepare the data for processing by machine learning algorithms. The preprocessing steps include:

1. Data Cleaning – Removing duplicate entries and handling missing values.
2. Feature Engineering – Transforming categorical attributes into numerical values using label encoding, as well as creating derived variables such as combinations of symptoms (e.g., "cough + fever").
3. Labeling – Defining the target variable (label) as "Bronchitis" (1) and "Non-Bronchitis" (0).

This preprocessing process is essential to ensure that the data used for model training are of high quality, free from noise, and compatible with the CatBoost input format.

2.3. Data Splitting

After preprocessing, the dataset is divided into three training and testing ratios: 60:40, 70:30, and 80:20. This step aims to observe how variations in the amount of training data affect the model's performance.

2.4. Hyperparameter Tuning

The parameter optimization process is carried out using the GridSearchCV technique, which is a method for identifying optimal hyperparameter values by exhaustively exploring predefined parameter combinations to achieve the best model performance.

2.5. Catboost Algorithm

CatBoost (Categorical Boosting) is a gradient boosting algorithm developed by Yandex Research and specifically designed to efficiently handle categorical data. CatBoost operates by sequentially combining multiple simple decision trees, where each new tree attempts to correct the prediction errors of the previous trees. This approach is based on the gradient descent principle, which iteratively minimizes the loss function, resulting in increasingly accurate predictions.

2.6. Model Evaluation

The evaluation stage aims to measure the model's ability to accurately predict indications of bronchitis. Model evaluation is performed using a Confusion Matrix and the ROC-AUC Curve as classification performance indicators.

The main evaluation metrics used include:

1. Accuracy, to measure the overall proportion of correct predictions.
2. Precision, to assess the correctness of positive (bronchitis) predictions.
3. Recall, to evaluate the model's ability to detect actual bronchitis cases.
4. F1-Score, which represents the harmonic mean of precision and recall.
5. ROC-AUC, to assess the model's capability to distinguish between positive and negative classes.

3. FINDINGS AND DISCUSSION

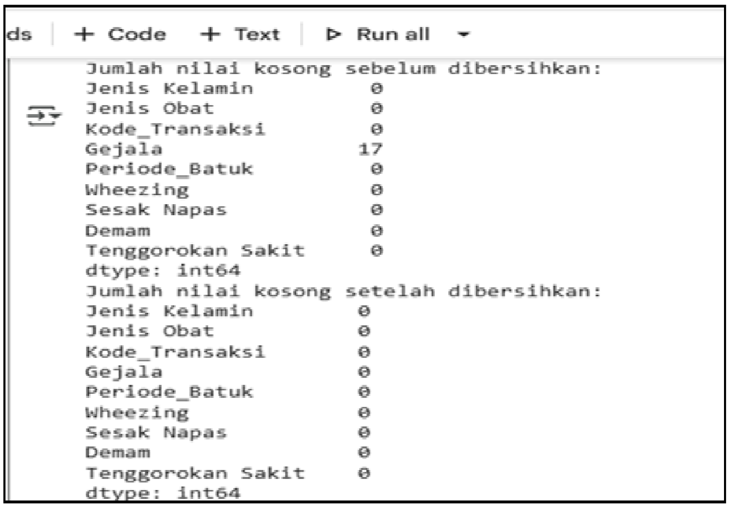
3.1 Data Collection

At the data collection stage, the dataset was collected during the period of January–March 2025. Patient data from RH Farma Pharmacy, stored in Excel format on Google Drive, were loaded into the program for further processing.

3.2 Data Cleaning

The data cleaning stage involved handling missing values, removing duplicate records, and eliminating irrelevant columns to ensure that the dataset was clean and consistent prior to modeling.

Figure 2. Datasets after cleaning



```
Jumlah nilai kosong sebelum dibersihkan:
Jenis Kelamin          0
Jenis Obat             0
Kode_Transaksi         0
Gejala                17
Periode_Batuk          0
Wheezing               0
Sesak Napas            0
Demam                  0
Tenggorokan Sakit      0
dtype: int64
Jumlah nilai kosong setelah dibersihkan:
Jenis Kelamin        0
Jenis Obat           0
Kode_Transaksi       0
Gejala               0
Periode_Batuk        0
Wheezing             0
Sesak Napas          0
Demam                0
Tenggorokan Sakit    0
dtype: int64
```

Figure 2 illustrates the number of missing values before and after the data cleaning process. Prior to cleaning, only the Symptoms column contained 17 missing values, while all other columns were complete. After the cleaning process, all columns—including the Symptoms column—contained no missing values, indicating that the dataset was ready for subsequent analysis.

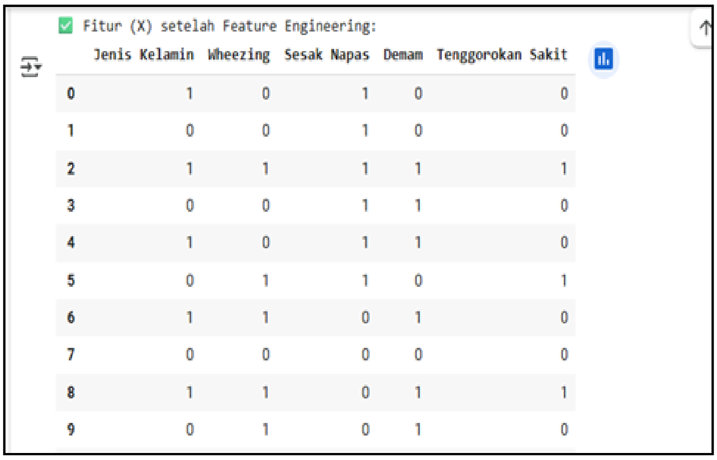
3.3 Labeling

The labeling process was conducted by adding a target variable, namely Bronchitis Indication. Label determination was based on the duration of the patient's cough. Patients who experienced coughing for more than three weeks were assigned a label of 1 (indicated bronchitis), whereas patients who coughed for three weeks or less, or had no cough symptoms, were assigned a label of 0 (non-bronchitis).

3.4 Feature Engineering

The feature engineering stage involved converting categorical features such as Gender, Wheezing, Shortness of Breath, Fever, and Sore Throat into numerical form. The value "yes" was encoded as 1 and "no" as 0, while Gender was encoded as 1 for male and 0 for female, enabling the data to be processed by the CatBoost algorithm.

Figure 3. Datas after Feature Engineering


✅ Fitur (X) setelah Feature Engineering:

	Jenis Kelamin	Wheezing	Sesak Napas	Demam	Tenggorokan Sakit
0	1	0	1	0	0
1	0	0	1	0	0
2	1	1	1	1	1
3	0	0	1	1	0
4	1	0	1	1	0
5	0	1	1	0	1
6	1	1	0	1	0
7	0	0	0	0	0
8	1	1	0	1	1
9	0	1	0	1	0

Figure 3 presents the final results of the categorical feature encoding process, where categorical values were transformed into numerical representations. After this process, five new columns were generated in the dataset, which were ready to be used by the model.

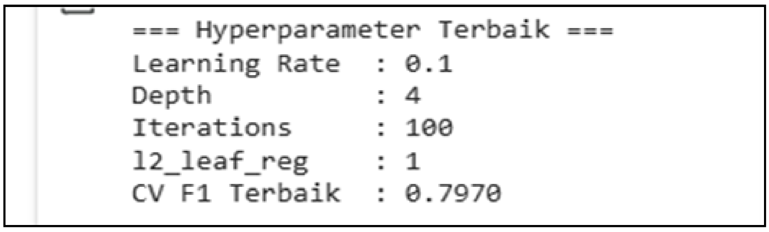
3.5 Data Splitting

The data splitting stage began with separating features and the target variable. The features used included Gender, Wheezing, Shortness of Breath, Fever, and Sore Throat, while the target variable was Bronchitis. The dataset was then divided into training and testing sets using *the train_test_split()* function with ratios of 60:40, 70:30, and 80:20. The stratify parameter was applied to maintain balanced class distribution, and the resulting splits were stored as dataframes for documentation purposes.

3.6 Hyperparameter Tuning

To obtain an optimal model configuration, hyperparameter tuning was performed on the CatBoost algorithm using the GridSearchCV method with cross-validation (Stratified K-Fold). This process aimed to identify the best combination of parameters such as learning rate, depth, iterations, and l2_leaf_reg to achieve maximum predictive performance.

Figure 4. Datas after Hyperparameter Tuning



```
=== Hyperparameter Terbaik ===
Learning Rate   : 0.1
Depth           : 4
Iterations      : 100
l2_leaf_reg     : 1
CV F1 Terbaik   : 0.7970
```

Based on Figure 4, the hyperparameter tuning results using GridSearchCV with Stratified K-Fold cross-validation produced the best parameters for the CatBoost algorithm: a learning rate of 0.1, depth of 4, iterations of 100, and l2_leaf_reg of 1. This parameter combination achieved an F1-Score of 0.7970, indicating a strong balance between precision and recall.

3.7 Model Implementation

After obtaining the optimal hyperparameter configuration, the next stage was model implementation using the CatBoost algorithm with the selected parameters (learning rate 0.1, depth 4, iterations 100, and l2_leaf_reg 1). The model was trained and evaluated under three data split scenarios (60:40, 70:30, and 80:20) to compare performance. Prediction results were evaluated using Accuracy, Precision, Recall, and F1-Score metrics to assess CatBoost’s effectiveness in detecting bronchitis indications.

Table 1. Catboost Algorithm Model Results

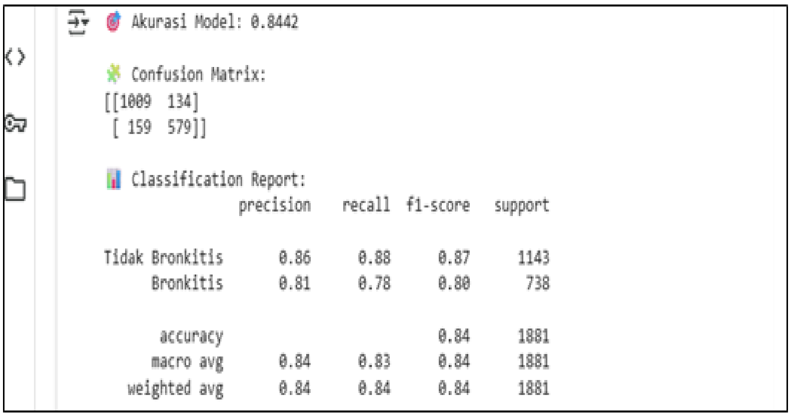
Data Split	*Accuracy*	*Precision*	*Recall*	*F1-Score*
60:40	0.8181	0.75232	0.7966	0.7743
70:30	0.8265	0.7925	0.7568	0.7742
80:20	0.8223	0.7914	0.7432	0.7666

Table 1 presents the performance results of the CatBoost model under three data split scenarios: 60:40, 70:30, and 80:20. Based on the evaluation, the 60:40 split achieved the best performance, with an accuracy of 0.8181 and the highest F1-Score of 0.7743, indicating a strong balance between precision (0.7523) and recall (0.7966). This ratio enabled the model to optimally capture data patterns while maintaining good generalization ability. Meanwhile, the 70:30 and 80:20 splits demonstrated relatively stable but slightly lower performance compared to the 60:40 split.

3.8 Model Evaluation

The final stage involved evaluating the CatBoost model with the optimal hyperparameter configuration to assess accuracy and predictive capability for bronchitis indication. The model with parameters (learning rate 0.1, depth 4, iterations 100, and l2_leaf_reg 1) demonstrated optimal performance based on Accuracy, Precision, Recall, and F1-Score metrics. The F1-Score was used as the primary evaluation metric, as it reflects the balance between precision and recall. These results indicate that CatBoost performs well in predicting bronchitis indications.

Figure 5. Model Evaluation



```
🎯 Akurasi Model: 0.8442

🧩 Confusion Matrix:
[[1009  134]
 [ 159  579]]

📊 Classification Report:
                 precision    recall  f1-score   support

Tidak Bronkitis       0.86      0.88      0.87      1143
      Bronkitis       0.81      0.78      0.80       738

       accuracy                           0.84      1881
      macro avg       0.84      0.83      0.84      1881
   weighted avg       0.84      0.84      0.84      1881
```

Figure 5 presents the results of the CatBoost model evaluation using the classification_report function. The model achieved an accuracy of 0.8442, or approximately 84%, indicating that the majority of predictions were consistent with the actual labels. Based on the confusion matrix, out of 1,143 non-bronchitis patients, 1,009 were correctly classified while 134 were misclassified. Meanwhile, out of 738 patients indicated with bronchitis, 579 were correctly detected and 159 were misclassified. These results demonstrate that the model exhibits reasonably strong classification performance in distinguishing between bronchitis and non-bronchitis patients.

4. CONCLUSION

Based on the results of the research and the evaluations conducted, it can be concluded that the CatBoost algorithm is capable of delivering strong performance in predicting indications of bronchitis based on patient symptom data obtained from RH Farma Pharmacy. The developed model is able to capture relevant patterns from patient symptoms and effectively classify cases of bronchitis and non-bronchitis with a high level of accuracy.

The experimental results indicate that the 60:40 data split scenario yields the best performance, achieving an accuracy of 0.8181 and the highest F1-Score of 0.7743, which reflects an optimal balance between precision and recall. The model with optimal parameters (learning rate of 0.1, depth of 4, iterations of 100, and l2_leaf_reg of 1) also achieves an overall accuracy of approximately 84%, indicating a reliable classification capability.

Overall, the CatBoost algorithm has proven to be effective and stable in detecting indications of bronchitis and demonstrates strong potential for further development to support data-driven disease prediction systems in the medical domain.

REFERENCES

Annisa, G., & Khairani, R. (2024). Smoking and Allergies as Factors Associated with Acute Bronchitis in Adult Patie~nts. Jurnal Akta Trimedika (JAT), 1(3), 316–326.

Hariningsih, S., Sujangi, & Prasetyo, A. (2023). The Influence of the Home Physical E~nvironment and Behavioral Factors on the Incidence of Acute Respiratory Tract Infections (ISPA). Gema Lingkungan Kesehatan, 21(2), 51–58.

Ilmi, M. B., & Kusrini. (2025). A Performance Comparison of Machine Learning Algorithms for Detecting Potential HIV Risk. J. Buffer Informatika, 11(1), 1–10.

Lekatompessy, P., & Magafira, P. (2023). Nursing Care for Patients with Bronchitis in the St. Bernadeth II Ward of Stella Maris Hospital, Makassar. Final Scientific Project, STIKes Stella Maris, Makassar.

Permatasari, N. L., Syahidah, S. A., Irfiansyah, A. L., & Al-Haqqoni, M. G. (2022). Predicting Diabetes Mellitus Using the CatBoost Classifier and the Shapley Additive Explanation (SHAP) Approach. Barekeng: Journal of Mathematics and Applied Sciences, 16(2), 615–624.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2020). CatBoost: Unbiased Boosting with Categorical Features. Advances in Neural Information Processing Systems, 31, 1–11.

Sabili, N. L., Umbara, F. R., & Melina. (2024). Classification of Diabetes Disease Using the Categorical Boosting Algorithm with Diabetes Risk Factors. JATI (Journal of Informatics Engineering Students), 8(6), 1–10.

Syamkalla, M. T., Khomsah, S., & Nur, Y. S. R. (2024). Implementation of the CatBoost Algorithm and Shapley Additive Explanations (SHAP) for Predicting the Popularity of Indie Games on the Steam Platform. Journal of Information Technology and Computer Science, 11(4), 777–786.

Ishlah, A. W., Sudarno, S., & Kartikasari, P. (2023). Implementation of GridSearchCV in Support Vector Regression (SVR) for Stock Price Forecasting. Gaussian Journal, 12(2), 276–286.

Chang, W., Wang, X., Yang, J., & Qin, T. (2023). An Improved CatBoost-Based Classification Model for Ecological Suitability Assessment of Blueberries. Sensors, 23(4), 1811.