



Perbandingan Algoritma CatBoost dan XGBoost dalam Klasifikasi Penyakit Jantung

Yoan Purbolingga*, Dila Marta Putria^a, Fahrizal^a, Asde Rahmawatia^a, Bastul Wajhi Akramunnas^a

^aTeknik Elektro, Institut Teknologi Bisnis Riau, Pekanbaru

INFO ARTIKEL

Histori artikel:
Tersedia Online: 24 Juni 2023

ABSTRAK

Penyakit jantung merupakan masalah kesehatan yang serius dan dapat berdampak negatif pada kualitas hidup individu. Dalam beberapa tahun terakhir, penggunaan teknik pembelajaran mesin dalam mendiagnosis penyakit jantung telah menjadi topik penelitian yang penting. Pada penelitian ini melakukan perbandingan kinerja algoritma CatBoost dan XGBoost, dalam tugas klasifikasi penyakit jantung. Tahapan perbandingan algoritma diawali dengan data acquisition, exploratory data analysis, future engineering, modelling, dan model evaluation. Hasilnya membuktikan bahwa algoritma CatBoost dapat mengungguli algoritma XGBoost dalam mengklasifikasi kasus orang dengan penyakit jantung dengan data set yang digunakan. Untuk dataset yang digunakan diperoleh dari UCI Machine Learning Repository dengan 12 feature dan 918 jumlah data. Dari hasil analisis data, feature OldPeak, MaxHR, Age, dan FastingBS merupakan faktor utama dalam memprediksi orang dengan penyakit jantung. Sehingga berdasarkan kinerja algoritma CatBoost dan XGBoost dengan data ini, kami merekomendasikan algoritma CatBoost untuk memprediksi orang dengan penyakit jantung dengan lebih baik.

Kata kunci: Cat Boost; XG Boost; Penyakit jantung

E - MAIL

Johanyurik@gmail.com*
putri.dilamarta@gmail.com
fahrizal.upp@gmail.com
iamtitiuin@gmail.com
bastulwajhiakramunnas@gmail.com

ABSTRACT

Heart disease is a serious health problem and can hurt an individual's quality of life. In recent years, the use of machine learning techniques in diagnosing heart disease has become an important research topic. In this study, we compared the performance of the CatBoost and XGBoost algorithms in the classification task of heart disease. The algorithm comparison stage begins with data acquisition, exploratory data analysis, future engineering, modeling, and model evaluation. The results prove that the CatBoost algorithm can outperform the XGBoost algorithm in classifying cases of people with heart disease with the data set used. The dataset used was obtained from the UCI Machine Learning Repository with 12 features and 918 amounts of data. From the results of data analysis, the OldPeak, MaxHR, Age, and FastingBS features are the main factors in predicting people with heart disease. So based on the performance of the CatBoost and XGBoost algorithms with this data, we recommend the CatBoost algorithm to predict people with heart disease better.

Keywords: Cat Boost; XG Boost; Heart disease

I. PENDAHULUAN

Di Indonesia, penyakit jantung masuk ke dalam salah satu penyakit yang paling banyak menyebabkan kematian. Jantung memiliki tugas penting dalam memompa darah yang membawa zat

oksigen dan nutrisi terhadap seluruh sel-sel tubuh untuk kelangsungan hidup mereka, serta membantu dalam menghilangkan metabolisme yang sudah tidak berguna dalam tubuh [1].

Penyakit jantung merupakan penyakit yang terjadi pada sistem pembuluh darah yang mengakibatkan ketidakseimbangan antara suplai dan kebutuhan darah. Penyakit jantung dan pembuluh darah semakin meningkat dari tahun ke tahun. Berdasarkan data Riset Kesehatan Dasar (Riskesdas) tahun 2018, 15 dari 1000 orang, atau sekitar 2.784.064 individu di Indonesia menderita penyakit jantung. Beberapa penyebab dari kelainan jantung ini, diantaranya adalah serangan jantung, tekanan darah tinggi, stres, Usia, kadar trigliserida, hipertensi, dan diabetes melitus [2].

Penyakit jantung memiliki beberapa jenis klasifikasi, diantaranya adalah *myocardial infarction*, dan *heart failure*. Klasifikasi penyakit jantung merupakan salah satu tugas penting dalam bidang medis untuk membantu dalam diagnosis dan pengobatan penyakit jantung. Dalam beberapa tahun terakhir, pengembangan dan penerapan teknik pembelajaran mesin dalam klasifikasi penyakit jantung telah menjadi area penelitian yang aktif. Seperti penggunaan teknik *machine learning*. Penggunaan *machine learning* terbukti telah banyak diterapkan dalam topik klasifikasi dan optimasi dalam membuat sistem cerdas untuk meningkatkan penyedia layanan kesehatan [3], algoritma *machine learning* ini juga dapat digunakan sebagai alat bantu melakukan diagnosis medis [4].

Algoritma *machine learning* merupakan salah satu dari perkembangan teknologi yang saat ini terus berkembang. Seiring berjalannya waktu perlu menciptakan metode yang terbaik untuk dapat mengolah data menjadi semakin efisien dan akurat terus dikembangkan dan diterapkan [5]. Maka perlu dilakukan perbandingan dari model-model yang ada sehingga mampu menghasilkan kinerja yang paling baik. Algoritma *CatBoost* dan *XGBoost* adalah jenis dari *ensemble technique* yang merupakan kembangan dari algoritma *machine learning*. Sehingga pada penelitian ini dilakukan analisa perbandingan model *CatBoost* serta *XGBoost* terhadap penggunaannya pada masalah klasifikasi kasus orang dengan penyakit jantung dengan data set yang tersedia.

Setiap model algoritma klasifikasi memiliki kelebihan dan kelemahan masing-masing dalam menangani klasifikasi dari suatu data set [6]. Hal ini terjadi karena setiap data set yang digunakan memiliki karakteristik yang berbeda juga. Analisa akan dilakukan dengan mempertimbangkan akurasi

dari kedua model. Sehingga akan diperoleh model yang paling baik.

II. MATERIAL DAN METODE

2.1 Extreme Gradient Boosting

Extreme Gradient Boosting atau lebih dikenal dengan (*XGBoost*) adalah model yang pertama kali diusulkan oleh Tianqi Chen dan Carlos Guestrin pada tahun 2011 dan terus dioptimalkan dan ditingkatkan dalam studi lanjutan dari banyak peneliti [7]. *XGBoost* merupakan salah satu dari *boosted tree algorithms*, yang mengikuti aturan *gradient boosting*.

Jika dibandingkan dengan algoritma *gradient boosting* lainnya, *XGBoost* menggunakan formalisasi model yang lebih teratur untuk mengontrol data *over-fitting*, sehingga memberikan kinerja lebih baik [8]. Yang perlu kita diketahui adalah fungsi f_t yang masing-masing memuat struktur *tree* dan *leaf score*. Ini dapat diformulasikan seperti yang terlihat pada (1):

$$f_t(x) = w_{q(x)}, w \in R^d \rightarrow \{1, 2, \dots, T\} \quad (1)$$

dimana w merupakan vector dari skor yang terdapat pada *leaves*, q merupakan fungsi yang menetapkan setiap titik data ke *leaf* yang sesuai dan T adalah jumlah *leaf*. Kompleksitas model diformulasikan sebagai:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (2)$$

Tujuan fungsi pada iterasi ke- t terlihat pada (3):

$$Obj^t = \sum_{j=1}^T \left[G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma T \quad (3)$$

Memecahkan *quadratic* ini (3), w_j terbaik untuk struktur yang diberikan $q(x)$ dan reduksi terbaik yang didapatkan adalah:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \quad (4)$$

$$Obj^* = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T \quad (5)$$

Skor yang diperoleh membagi *leaf* menjadi 2 adalah seperti yang terlihat pada (6):

$$Gain = \frac{1}{2} \left[\frac{GL^2}{HL+\lambda} + \frac{GR^2}{HR+\lambda} - \frac{(GL+GR)^2}{HL+HR+\lambda} \right] - \gamma \quad (6)$$

dimana:

$$G_j = \sum_{i \in I_j} g_i \quad \text{dan} \quad H_j = \sum_{i \in I_j} h_i$$

2.2 CatBoost

CatBoost atau *categorical boosting* adalah algoritma yang dikembangkan oleh Yandex. *CatBoost* merupakan salah satu algoritma yang mengimplementasikan *gradient boosting* dimana menggunakan *binary decision tree* sebagai *base predictors* [9]. *CatBoost* dapat menangani *features* kategorikal, *ordered* dan *overfitting* semuanya ditangani oleh *Bayesian estimators*. Algoritma *CatBoost* menggunakan *Prediction Values Change* (PVC) atau *Loss Function Change* (LFC) untuk menentukan peringkat *features* model yang dikembangkan. PVC adalah metode *default* yang digunakan dalam model *machine learning CatBoost*. LFC umumnya digunakan untuk memberi peringkat model tertentu di antara berbagai model [10].

$$F = \{f_1, f_2, f_3, \dots, f_n\} \quad (7)$$

$$P_i = \beta F_j \quad (8)$$

F merepresentasikan sekumpulan *features input*, β merupakan faktor numerik yang ditetapkan untuk *features input*, dan P sebagai prediksi langkah tertentu. P_i merepresentasikan nilai prediksi pada faktor numerik tersubstitusi, β merepresentasikan faktor numerik, dan F_j merepresentasikan *features* tertentu yang dipilih dari *set features* yang diberikan.

$$P_{i+1} = \beta_{i+1} F_j \quad (9)$$

P_{i+1} menyatakan nilai prediksi Ketika faktor numerik diubah, β_{i+1} menyatakan faktor numerik yang dimodifikasi.

$$P_{i=0} \neq P_i \neq P_{i+1} \quad (10)$$

menyatakan bahwa ketika ada perubahan dalam faktor numerik akan mengubah nilai prediksi, sehingga *features* tertentu diperlukan.

2.3 Training dan Testing Data set

Data set yang digunakan diperoleh dari UCI *Machine Learning Repository* yang terdiri dari 11 *feature* independen dan 1 *feature* dependen dengan 918 jumlah data. Rasio data testing sebesar 30%, data *training* sebesar 70% dan dengan pemilihan acak (*random*).

2.4 Metrik Evaluasi

Untuk menemukan algoritma terbaik, kami melatih model *CatBoost* dan *XGBoost* pada kumpulan data. Model klasifikasi ini akan dievaluasi menggunakan metrik kinerja berikut:

- *Accuracy*- rasio prediksi *True* (positif dan negatif) dengan keseluruhan data.
- *Precision*- rasio prediksi *True* positif dibandingkan dengan keseluruhan hasil yang diprediksi positif.
- *Recall*- rasio prediksi *True* positif dibandingkan dengan keseluruhan data yang benar positif.
- *Specificity*- kebenaran memprediksi negatif dibandingkan dengan keseluruhan data negatif.
- *F1 Score*- perbandingan rata-rata *precision* dan *recall* yang dibobotkan.

Dalam domain medis, *False Positive* (orang dengan resiko gagal jantung rendah atau tanpa risiko) tidak lebih berbahaya seperti *False negative* (orang dengan resiko gagal jantung tinggi tetapi tetapi tidak terdeteksi).

III. HASIL DAN PEMBAHASAN

Pada sub bab ini menjelaskan segala informasi yang telah didapatkan selama penelitian. Mulai dari tahap *data acquisition*, *exploratory data analysis*, *future engineering*, *modelling*, dan *model evaluation*.

3.1 Data Acquisition

Pada penelitian ini data set diperoleh dari UCI *Machine Learning Repository*, yang terdiri dari 11 *feature* independen dan 1 *feature* dependen dengan 918 jumlah data. Data set ini diambil dari 5 sumber yang berbeda untuk menciptakan data set yang lebih bervariasi. Detail data set dapat dilihat pada Tabel 1 dan 2.

Tabel 1. Sumber Data Set

Sumber	Jumlah
Cleveland	303
Hungarian	294
Switzerland	123
Long Beach VA	200
Stalog (<i>Heart</i>) Data Set	270
Total	1190
Duplikasi	272
Total Akhir	918

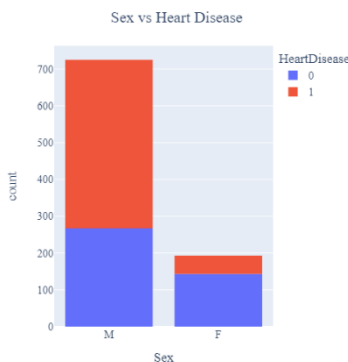
Tabel 2. Keterangan Data Set

Atribut	Deskripsi
Age	Umur (Continuous)
Sex	Gender (Nominal)
ChestPain	Jenis sakit dada (Nominal)
RestingBP	Tekanan darah (Continuous)
Cholesterol	Kolesterol (Continuous)
FastingBP	Tingkat gula darah (Nominal)
RestingECG	Efektifitas listrik di jantung (Nominal)
MaxHR	Detak jantung maksimum (Continuous)
ExerciseAngina	Latihan pada angina (Nominal)
Oldpeak	Depresi ST yang diakibatkan oleh latihan relatif saat istirahat (Continuous)
ST_Slope	Kemiringan segmen ST Latihan punjak (Nominal)
HeartDisease	Pembagian kelas penyakit gantung (Nominal)

3.2 Exploratory Data Analysis

Exploratory data analysis pada atribut yang disebutkan di atas dan memperoleh *insights* yang sangat berarti mengenai dampak atribut ini terhadap penyakit jantung.

1. Sex

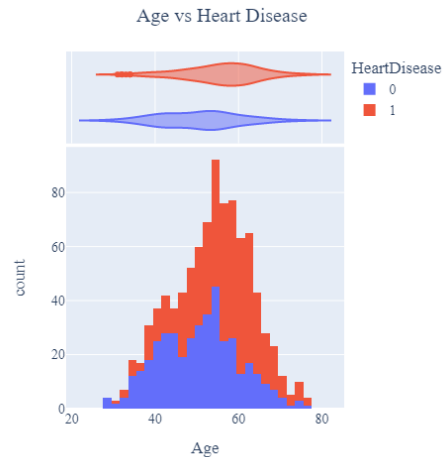


Gambar. 1. Korelasi antara *feature* Sex dengan HeartDisease

Dapat diketahui dari Gambar 1 bahwa jenis kelamin laki-laki lebih banyak terkena penyakit jantung dibandingkan Wanita. Hal ini Hal ini

menunjukkan bahwa laki-laki lebih berisiko terkena penyakit jantung dibandingkan wanita.

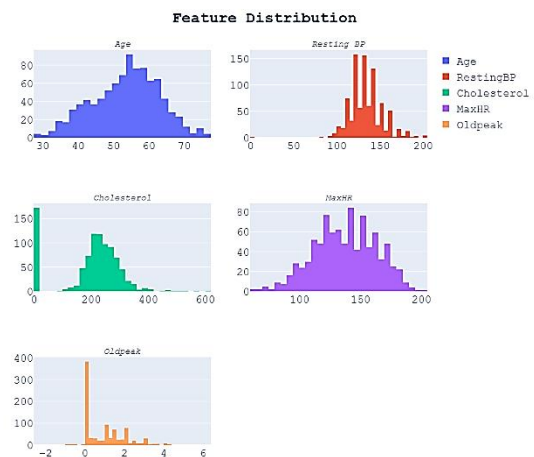
2. Age



Gambar. 2. Korelasi Antara *feature* Age dengan HeartDisease

Dari Gambar 2 dapat diketahui bahwa distribusi data orang dengan penyakit jantung banyak pada kisaran umur 54-55 tahun dengan jumlah 47 orang. Dari distribusi Gambar 2 juga dapat diketahui bahwa orang dengan umur kurang dari 30 tahun lebih sedikit terkena penyakit jantung dan memiliki ketahanan imun yang lebih tinggi dibandingkan orang yang sudah tua.

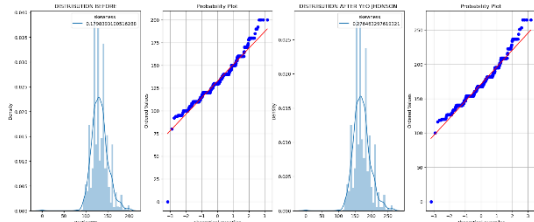
3. Features Numerical



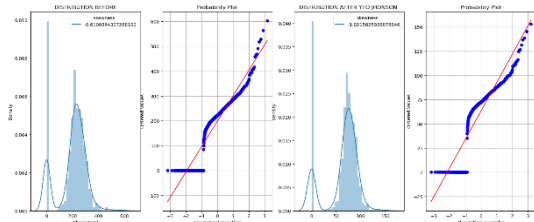
Gambar. 3. Distribusi *Features Numerical* pada Data Set

Pada Gambar 3 dapat dilihat data *numerical* atau *continuous* dan mendapati bahwa terdapat 3 *features* dengan tingkat *skewness* yang tinggi, yaitu *features* RestingBP, Cholesterol dan OldPeak. Sehingga dicoba penggunaan metode *yeo-johnson-transformations*.

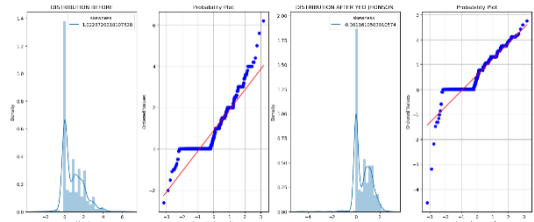
Dapat dilihat bahwa data banyak *skewness* ke kiri yang berarti negatif. Sehingga kita perlu menormalisasikannya dengan menggunakan metode *yeo-johnson-transformations* untuk *features* RestingBP, Cholesterol dan OldPeak.



Gambar. 4. Distribusi *Feature* RestingBP Sebelum dan Sesudah menggunakan Metode *Yeo-Johnson-Transformations*

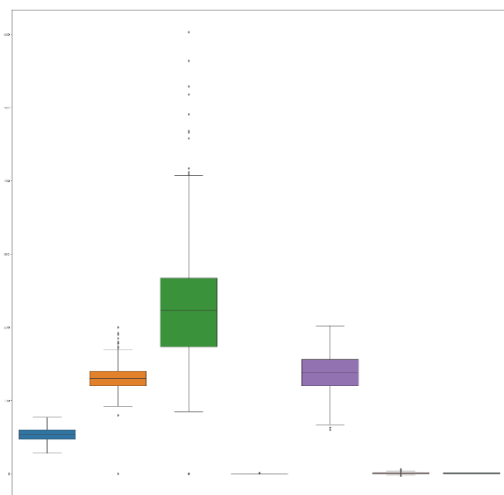


Gambar. 5. Distribusi *Feature* Cholesterol Sebelum dan Sesudah menggunakan Metode *Yeo-Johnson-Transformations*



Gambar. 6. Distribusi *Feature* OldPeak Sebelum dan Sesudah menggunakan Metode *Yeo-Johnson-Transformations*

4. *Outlier*

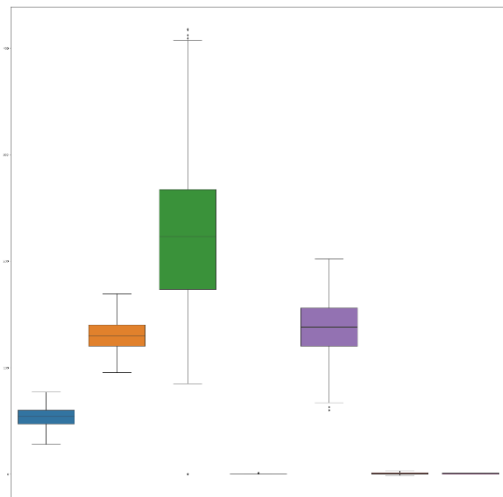


Gambar. 7. Boxplot Data *Outlier*

Column Age outliers = 0 => 0.0%
 Column RestingBP outliers = 42 => 4.575%
 Column Cholesterol outliers = 183 => 19.935%
 Column MaxHR outliers = 2 => 0.218%
 Column Oldpeak outliers = 16 => 1.743%

Gambar. 8. Banyak Data *Outlier*

Pada Gambar 7 menunjukkan bahwa ada banyak *features* yang memiliki *outlier*. Sehingga harus dibersihkan terlebih dahulu. Pembersihan data *outlier* menggunakan metode *z-score* dengan melihat distribusi datanya terlebih dahulu. Pembersihan dilakukan pada 3 *features* dengan jumlah *outlier* terbanyak yaitu *features* RestingBP, Cholesterol dan OldPeak. Sehingga setelah dilakukan proses pembersihan didapatkan data pada Gambar 9.



Gambar. 9. Boxplot Data *Outlier* (diterapkan metode *z score*)

5. *Korelasi antar Data Independen dan Dependen*



Gambar. 10. *Heatmap* data independen terhadap data dependen

Pada Gambar 10 dapat dilihat *heatmap* korelasi antara data independen dan data dependen. Disini data dependen merupakan *feature target* yaitu *feature* HeartDisease. Dari *heatmap* ini dapat

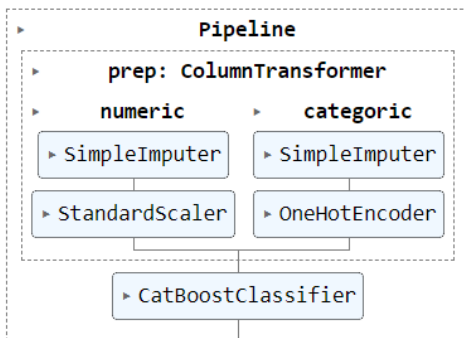
diketahui pula *feature importance* dari dataset ini adalah *feature* OldPeak, MaxHR, Age, FastingBS, Cholesterol dan RestingBP secara berurutan mulai dari yang terbesar hingga terkecil.

3.3 Feature Engineering

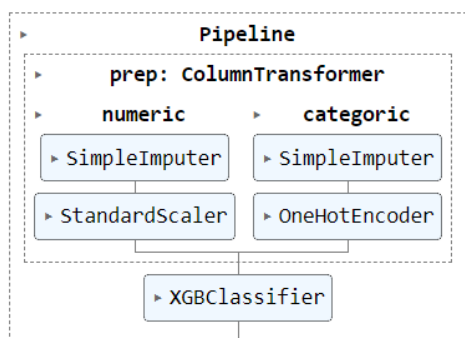
Pada tahap ini dilakukannya pemisahan antara data kategorikal (nominal) dengan data numerikal (*continuous*). Hal ini dilakukan agar diterapkannya metode yang sesuai untuk tahap *modelling* selanjutnya. Pada data kategorikal diterapkan “*one hot encoding*” sedangkan data *numerical* diterapkan “*standard scaler*”. Selain itu diberikan “*imputer*” untuk mencegah apabila terdapat data *null*.

3.4 Modelling

Dilakukan *modelling* menggunakan algoritma yang ingin diteliti yaitu *CatBoost* dan *XGBoost*.



Gambar. 11. Pipeline Model *CatBoost*



Gambar. 12. Pipeline Model *XGBoost*

Dari hasil *modelling* didapatkan hasil akurasi sebagai berikut:

Tabel 3. Hasil Akurasi Modelling

Model	Catboost	XGBoost
Accuracy	0.89130	0.844203
Precision	0.913580	0.895425
Recall	0.902439	0.835366
Specificity	0.875000	0.857143
F1 Score	0.907975	0.864353

Selain itu kami mencoba untuk melakukan *hyperparameter tuning* untuk kedua model dengan menambahkan parameter *learning_rate* = [0.01, 0.05, 0.1] dan menambahkan *Grid Search CV* dengan banyak 4 fold. *Grid Search CV* sendiri merupakan metode pemilihan kombinasi model dan *hyperparameter* dengan cara menguji coba satu persatu kombinasi dan melakukan validasi untuk setiap kombinasi. Tujuannya adalah untuk menentukan kombinasi yang menghasilkan performa model terbaik yang dapat dipilih untuk dijadikan model untuk prediksi. Dari uji coba *hyperparameter tuning* didapatkan hasil akurasi yang dapat dilihat pada Tabel 4.

Tabel 4. Hasil Akurasi *Hyperparameter Tuning*

Model	Catboost	XGBoost
Accuracy	0.887681	0.855072
Precision	0.913043	0.897436
Recall	0.896341	0.853659
Specificity	0.875000	0.857143
F1 Score	0.904615	0.875000

3.5 Evaluation Model

Melakukan model evaluation dengan melakukan oversampling dengan menggunakan metode SMOTE pada *data training* sehingga data menjadi seimbang.

Before OverSampling, counts of label '1': 344
Before OverSampling, counts of label '0': 298

Gambar. 13. Jumlah *data training* sebelum *Oversampling*

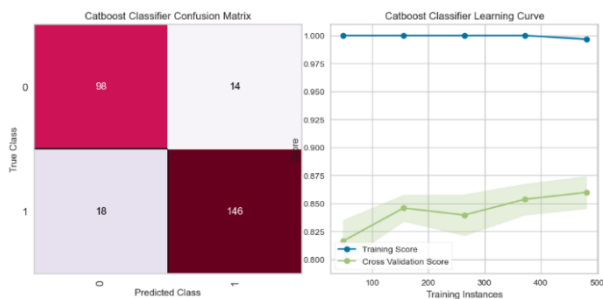
After OverSampling, counts of label '1': 344
After OverSampling, counts of label '0': 344

Gambar. 14. Jumlah *data training* setelah *Oversampling*

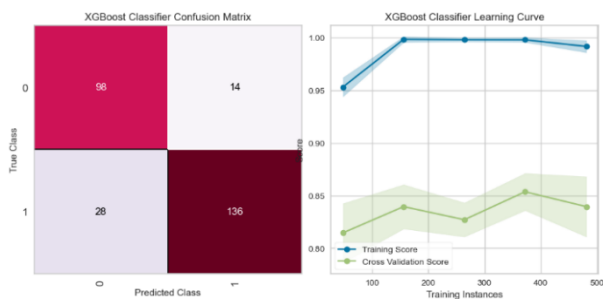
Dari hasil *evaluation model* didapatkan hasil akurasi masing-masing untuk algoritma *CatBoost* dan *XGBoost* sebagai berikut:

Tabel 5. Hasil Akurasi *Evaluation Model*

Model	Catboost	XGBoost
Accuracy	0.884058	0.847826
Precision	0.912500	0.906667
Recall	0.890244	0.829268
Specificity	0.875000	0.875000
F1 Score	0.901235	0.866242

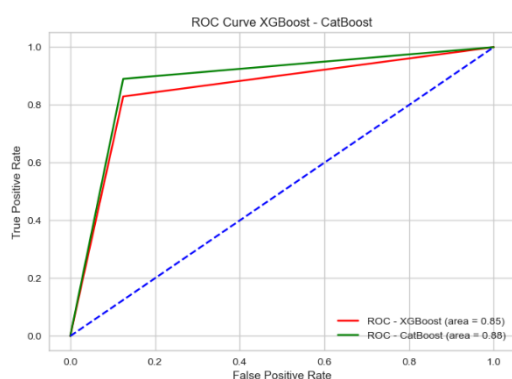


Gambar. 15. Confusion Matrix dan Learning Curve Model CatBoost



Gambar. 16. Confusion matrix dan learning curve model XGBoost

Gambar 15 dan 16 menunjukkan *confusion matrix* dari model *CatBoost* dan *XGBoost* yang menampilkan jumlah prediksi yang benar dan salah pada data uji dan menampilkan *learning curve* data *training* dan *testing* dari kedua model.



Gambar. 17. ROC-AUC Curve model CatBoost dan XGBoost

Gambar 17 adalah *ROC-AUC curve* yang merupakan salah satu metrik evaluasi terpenting untuk memeriksa kinerja model klasifikasi.

IV. KESIMPULAN

Jurnal ini mengusulkan model *ensemble learning-based* yang dibangun menggunakan dua model *machine learning* dengan tujuan memprediksi orang dengan risiko penyakit jantung. Kami menggunakan data set yang diperoleh dari *UCI Machine Learning Repository*, yang terdiri

dari 11 *feature* independen dan 1 *feature* dependen dengan 918 jumlah data. Setiap analisis kami melakukan *exploratory data analysis* dengan tujuan membersihkan data, mentransformasikan data dan mendapatkan insight dari data.

Kami mengimplementasikan dan melakukan analisis terhadap algoritma *CatBoost* dan *XGBoost*. Beberapa matriks evaluasi digunakan dalam jurnal ini untuk mengevaluasi performa model yang diimplementasikan.

Model *CatBoost* mendapatkan skor yang cukup baik dan berhasil mencapai skor tertinggi (akurasi). Dari hasil *hyperparameter tuning* dan evaluasi model juga menunjukkan bahwa model *CatBoost* mendapatkan skor lebih tinggi dibandingkan model *XGBoost*. Oleh karena itu, kami merekomendasikan model *CatBoost* sebagai model untuk memprediksi pada data orang dengan risiko penyakit jantung.

DAFTAR PUSTAKA

- [1] Rokom, "Penyakit Jantung Penyebab Utama Kematian, Kemenkes Perkuat Layanan Primer," 2022. <https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20220929/0541166/penyakit-jantung-penyebab-utama-kematian-kemenkes-perkuat-layanan-primer/>.
- [2] E. Margarini and M. Anindita, "Peringatan Hari Jantung Sedunia 2021: Jaga Jantungmu untuk Hidup Lebih Sehat," 2021. <https://promkes.kemkes.go.id/peringatan-hari-jantung-sedunia-2021-jaga-jantungmu-untuk-hidup-lebih-sehat>.
- [3] A. Byna and M. Basit, "Penerapan Metode Adaboost untuk Mengoptimasi Prediksi Penyakit Stroke dengan Algoritma Naïve Bayes," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 9, no. 3, 2020, doi: 10.32736/sisfokom.v9i3.1023.
- [4] R. Ahsana, R. R. Saedudin, and V. P. Widartha, "Perbandingan Akurasi Algoritma Adaboost and Algoritma Lightbm untuk Klasifikasi Penyakit Diabetes," *e-Proceeding Eng. Univ. Telkom Bandung*, vol. 8, no. 5, p. 9738, 2021.
- [5] S. Lukas, O. Vigo, D. Krisnadi, and P. Widjaja, "Perbandingan Performa Bagging dan Adaboost untuk Klasifikasi Data Multi-Class," *J. Inf. Syst. Dev.*, vol. 7, no. 2, 2022, doi: 10.19166/isd.v7i2.547.
- [6] Y. Pristyanto, "Penerapan Metode Ensembl untuk Meningkatkan Kinerja Algoritma Klasifikasi pada Imbalanced Dataset," *J. Teknoinfo*, vol. 13, no. 1, 2019, doi:

10.33365/jti.v13i1.184.

- [7] T. Chen and C. Guestrin, "XGBoost : Reliable Large-scale Tree Boosting System," *arXiv*, 2016.
- [8] R. Punnoose and P. Ajit, "Prediction of Employee Turnover in Organizations using Machine Learning Algorithms A case for Extreme Gradient Boosting," *Int. J. Adv. Res. Artif. Intell.*, vol. 5, no. 9, 2016.
- [9] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, 2018, vol. 2018-December.
- [10] B. Dhananjay and J. Sivaraman, "Analysis and classification of heart rate using CatBoost feature ranking model," *Biomed. Signal Process. Control*, vol. 68, 2021, doi: 10.1016/j.bspc.2021.102610.