# MACHINE LEARNING UNTUK PERBANDINGAN TINGKAT AKURASI PREDIKSI PENYAKIT DIABETES METODE SUPERVISED LEARNING

Ali Murtadho

Teknik Informatika , Fakultas Teknik, Universitas 17 Agutus 1945 Surabaya

Email: id.alimurtadho@gmail.com

Dwi Harini Sulistyawati

Teknik Informatika , Fakultas Teknik, Universitas 17 Agutus 1945 Surabaya

Email: Dwiharini@untag-sby.ac.id

## Abstract

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning (ML) techniques allows us to obtain predictive, the dataset we are testing is pima-indian-diabetes with a dataset of 768 raw data with 8 data features (Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI (Body Mass Index), Diabetes Pedigree Function, Age) and one dataset label (Outcome), we developed a method to achieve the best accuracy from the five methods we use with the stages of separation traning and testing the dataset, scaling features, parameters evaluation, confusion matrix and we get the accuracy of each method, and the results of the accuracy we get with these 5 methods *Gradient-boosting* is best with an accuracy score of 0.8, *Decision Tree* 0.72, *Random Forest* 0.72, next is *Logistic Regression* 0.7, and then followed by *K-NN* method with a score of 0.65.

**Keywords:** Machine Learning Prediction Diabetes, Performa Accuration Method , Supervised Learning, AI(artificial intelligence)

## Abstrak

Machine learning adalah aplikasi artificial intelligence (AI) yang menyediakan kemampuan pada sistem untuk secara otomatis belajar dan meningkatkan dari pengalaman tanpa diprogram secara eksplisit. Teknik Machine learning (ML) memungkinkan untuk mendapatkan hasil prediktif, dataset yang di buat uji coba adalah pima-indian-diabetes dengan dataset 768 data

mentah dengan delapan data fitur yang pertama Pregnancies, kedua Glucose, ketiga BloodPressure, keempat SkinThickness, kelima Insulin, keenam BMI (Body Mass Index), ke tujuh Diabetes Pedigree Function, kedelapan Age) dan satu label data (Outcome), disini teknik yang digunakan untuk mengembangkan metode mencapai akurasi terbaik dari ke lima metode yang di gunakan dengan tahapan pemisahan data traning dan pengujian dataset, Scaling feature, evaluasi parameter, Confusion matrik dan hasil uji coba dari akurasi masing-masing metode mendapatkan hasil sebgai berikut, Gradient boosting ini adalah metode yang terbaik dengan skor akurasi 0,8, Decision Tree 0,72, Random Forest 0,72, selanjutnya adalah Logistic regresion 0,7, kemudian diikuti oleh metode K-NN dengan skor 0,65.

.

Kata Kunci : Machine Learning Prediction Diabetes, Performa Accuration Method , Supervised Learning, AI(artificial intelligence).

## 1. PENDAHULUAN

Diabetes adalah penyakit jangka panjang atau kronis dan ditandai oleh kadar gula darah tinggi (glukosa) atau di atas nilai normal. Glukosa yang menumpuk di dalam darah karena tidak diserap oleh sel-sel tubuh dengan baik dapat menyebabkan berbagai gangguan pada organ tubuh. Jika diabetes tidak terkontrol dengan baik, berbagai komplikasi yang dapat membahayakan nyawa pasien dapat muncul [1]. Machine Learning (ML) adalah salah satu cabang dari disiplin Artificial Intelligence (AI) yang membahas pengembangan sistem berdasarkan data. Banyak hal yang dipelajari, tetapi pada dasarnya ada 4 hal utama yang dipelajari dalam pembelajaran mesin [2].

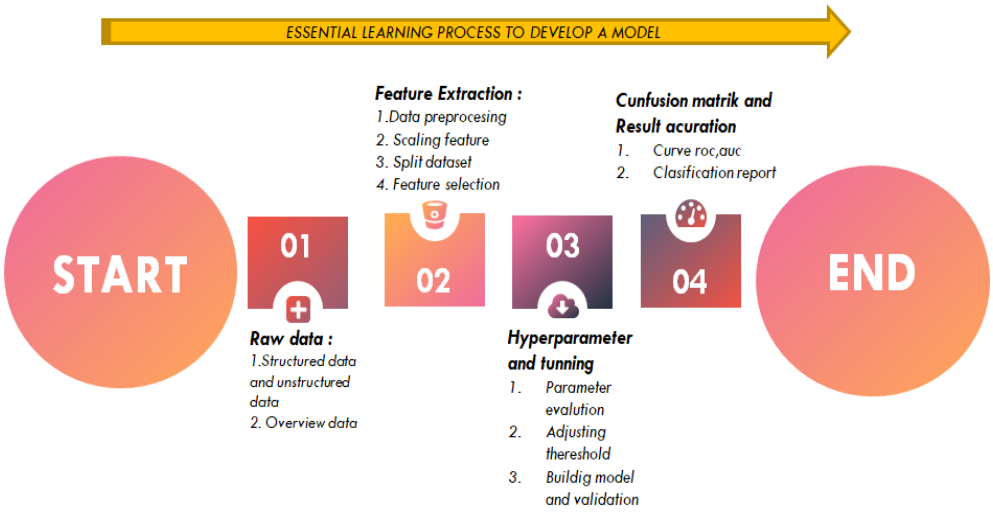
1. Supervised Learning.
2. Unsupervised Learning.
3. Semi-Supervised Learning.
4. Reinforcement Learning [3].

Teknik klasifikasi dalam penelitian ini menghasilkan model prediksi yang lebih akurat seperti itu adalah salah satu teknik penerapan pembelajaran Machine Learning (ML) yang paling umum melatih data dan membuat fungsi disimpulkan, yang dapat digunakan untuk memetakan contoh baru atau tidak terlihat. Tujuan utama dari teknik klasifikasi adalah untuk secara akurat memperkirakan kelas target untuk setiap kasus dalam data. Algoritma klasifikasi umumnya mensyaratkan kelas didefinisikan berdasarkan nilai atribut data.

Dataset yang kami uji kali ini kami ambil dari pima indian diabetes spesifikasi fitur 8 data mentah dan label data, jumlah data yang kami uji berjumlah 785 dataset, untuk proses pembelajaran saat ini kami menerapkan beberapa tahap untuk melihat kinerja akurasi dari metode kita akan menguji langkah-langkah yang kita gunakan [4].

Percobaan ini pertama Data mentah, kedua Ekstraksi fitur, tahap ketiga Hyperparameter dan Tunning, keempat Confusion Matriks dan hasil akurasi. Untuk lebih jelasnya bisa dilihat pada Gambar 1. Dalam penelitian ini, kami telah mempelajari kinerja lima model yang

berbeda untuk membandingkan akurasi model, model pertama Gradient Boosting, model kedua K-Nearest Neighbor (K-NN), model ketiga Decision Tree, model keempat Regresi Logistik, model kelima Random Forest [5].




*Gambar 1*

## 2. METODE PENELITIAN

### 2.1. Metodologi

Dalam model ini, kami telah menggunakan lima metode;

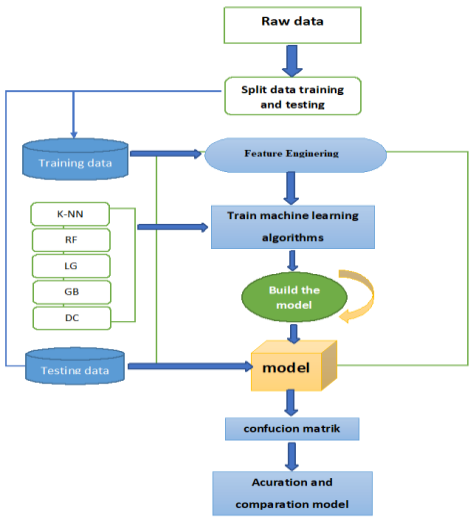
1. Gradient Boosting,
2. Random Forest,
3. Decision Tree,
4. Regresi Logistik,
5. K-NN

Untuk perbandingan akurasi, fokus utama dalam penelitian ini adalah mengetahui seberapa akurat metode yang akan di gunakan sehingga dapat mengetahui hasil dari metode mana yang akan menghasilkan akurasi terbaik dan mana yang merupakan akurasi terburuk di antara lima metode. Untuk dataset ada total 785 baris dan dibagi menjadi dua kelas: penderita diabetes dan non-penderita diabetes dengan delapan data fitur, delapan fitur adalah

1. Kehamilan,
2. Glukosa,
3. Tekanan Darah,
4. Tekanan Darah,
5. Insulin,
6. BMI (Indeks Massa Tubuh),

7. Fungsi Silsilah Diabetes,

8. Usia, dan kami menyertakan aliran proses kinerja algoritma dalam proses pembelajaran seperti yang ditunjukkan pada Gambar 3.




*Gambar 2 flow process*

Kumpulan data ini diperoleh dari Repositori UCI dari Database Pembelajaran Mesin.Kumpulan data dipilih dari kumpulan data yang lebih besar yang dipegang oleh Institut Nasional Diabetes dan Penyakit Pencernaan dan Ginjal. Semua pasien dalam database ini adalah wanita Pima-India yang berusia setidaknya 21 tahun dan tinggal di dekat Phoenix, Arizona, AS. Variabel respon biner mengambil nilai '0' atau '1', di mana '1 means berarti tes positif untuk diabetes dan' 0 is adalah tes negatif untuk diabetes. Ada 268 (34,9%) kasus di kelas '1' dan 500 (65,1%) kasus di kelas '0'.

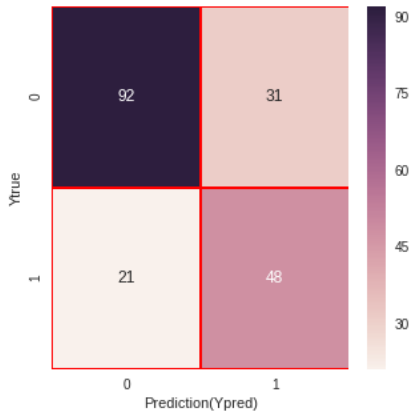
```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
Pregnancies                 768 non-null int64
Glucose                     768 non-null int64
BloodPressure               768 non-null int64
SkinThickness               768 non-null int64
Insulin                     768 non-null int64
BMI                         768 non-null float64
DiabetesPedigreeFunction    768 non-null float64
Age                         768 non-null int64
Outcome                     768 non-null int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

*Gambar 3 info dataset*

## 2.2. Gradient Boosting

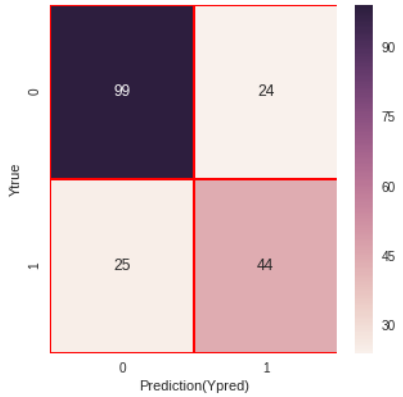
Dalam algoritma gradien boosting kali ini kami menggunakan ambang pengembangan penyesuaian Gradient Boosting Classifier (fitur pertama learning_rate dengan nilai= 0,05,fitur kedua max_depth dengan nilai= 3,fitur ketiga max_features dengan nilai= 0,5, random_state = 42),dan menghasilkan akurasi ini pada dataset pelatihan: 0,882, akurasi pada dataset testing: 0,750 dan dapatkan hasil kebingungan matriks seperti pada Gambar 4.




*Gambar 4 confusion matrix gradient boosting*

### 2.3. Decision Tree

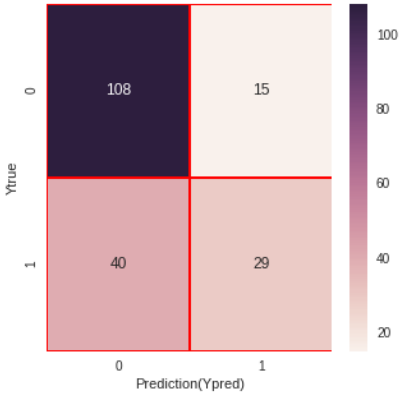
Algoritma Decision Tree yang kita gunakan untuk menetukan ambang batas parameter fitur algoritma Decision Tree Classifier adalah (yang pertama max_depth = dengen nilai 6, yang kedua max_features = dengan nilai 4, ketiga min_samples_split = dengan nilai 4, dan random_state = 42), menghasilkan akurasi pada dataset pelatihan ini: 0.852, akurasi pada dataset testing: 0.729 dan mendapatkan hasil dari matriks kebingungan seperti yang ditunjukkan pada Gambar 6.




*Gambar 5 confusion matriks decision tree*

### 2.4. K-NN (Kneighbors Classifier)

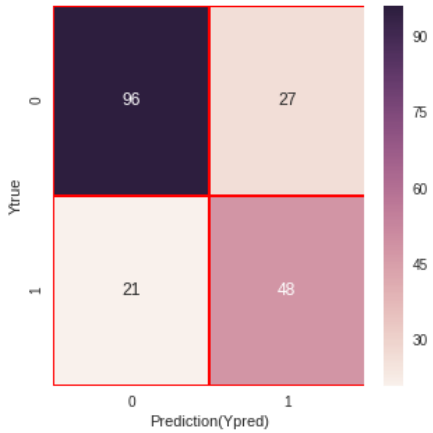
Algoritma K-NN menggunakan fitur ambang batas penyesuaian Kneighbors Classifier dimana di dalam fitur neigbors terdapat beberapa parameter di dalamnya di antaranya adalah (fitur pertama variabel algoritma dengan parameter = 'otomatis', fitur kedua leaf_size dengan nilai = 30,fitur ketiga metrik dengan parameter = 'minkowski', fitur keempat metric_params = Tidak ada, fitur kelima n_jobs dengan nilai = 1,fitur keenam n_neighbors dengan nilai = 5,fitur ketujuh p dengan nilai= 2,fitur kedelapan bobot dengan nilai = 'seragam' ), dan menghasilkan akurasi pada dataset pelatihan ini: 0,77, akurasi pada dataset testing: 0,71 dan dapatkan hasil dari matriks kebingungan seperti yang ditunjukkan pada Gambar 5.




*Gambar 6 confusion matriks K-NN*

### 2.5. Random Forest

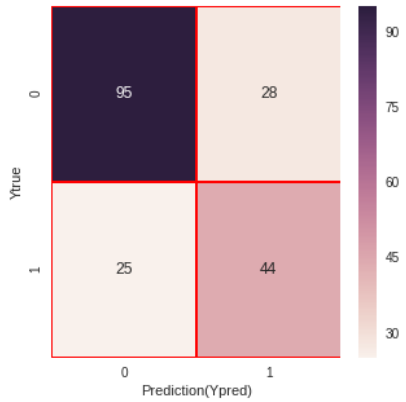
Algoritme Random Forest menggunakan ambang fitur penyesuaian Kelas diantaranya adalah (fitur yang pertama n_estimators dengan nilai = 100,fitur kedua kriteria dengan parameter = 'gini',fitur ketiga max_depth dengan nilai = 6,fitur keempat max_features dengan parameter = 'auto', random_state = 0),dan menghasilkan akurasi pada dataset pelatihan sebesar: 0,917, akurasi pada datset testing: 0,745 dan dapatkan hasil dari kebingungan matriks seperti yang ditunjukkan pada Gambar 7.




*Gambar 7 confusion matrik Random Forest*

2.6. Logistic Regresion

Algoritma Logistic Regresion menggunakan ambang batas fitur menyesuaikan fitur dari C dengan parameter logreg_classifier = Regresi Logistik (fitur pertama C dengan nilai = 1, fitur kedua penalti dengan nilai = 'l1'), dan menghasilkan akurasi pada dataset pelatihan : 0,783, akurasi pada dataset testing: 0,724 mendapatkan hasil dari matriks kebingungan sebagai ditunjukkan pada Gambar 7.




*Gambar 8 confusion matrik logistic regersion*

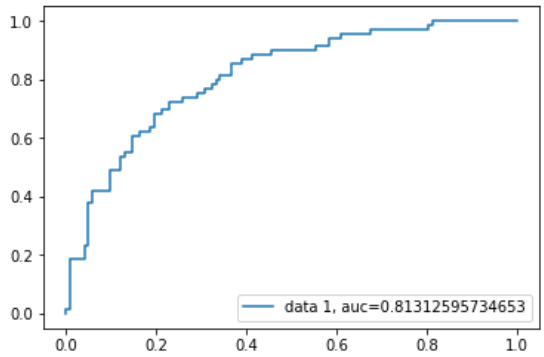
3. HASIL DAN PEMBAHASAN

Hasil akhir dari percobaan dari confusion matriks tiap-tiap algoritma akan di teruskan dengan perhitungan classification report yang meliputi [6].:

1. Precision = (TP / (TP + FP)) * 100%.
2. Recall = (TP / (TP + FN)) * 100%.
3. F1 Score = 2 * (Recall*Precission) / (Recall + Precission)
4. Akurasi = (TP + TN) / (TP + TN + FP + FN)) * 100%
5. Macro avg = menghitung metrik secara independen untuk setiap kelas dan kemudian mengambil rata-rata (karenanya memperlakukan semua kelas secara merata).
6. Weighted avg = mengembalikan rata-rata dengan mempertimbangkan proporsi untuk setiap label dalam dataset

Dari hasil classification report tersebut tiap metode akan mendapatkan hasi akurasi terbaik dan akan di jelaskan pada curva roc_auc pada gambar di bawah ini.

3.1.Gradient boosting




*Gambar 9 curva roc_auc algoritma gradient boosting*

Gambar 9 adalah preview hasil akurasi metode Gradient Boosting. Grafik yang dijelaskan pada gambar 9 adalah grafik liniear dengn warna biru untuk varibel horizontal adalah range angka terkecil akurasi sampai terbesar dan untuk variabel vertikal sama berarti range angka terkecil akurasi terkecil sampai terbesar dengan perhitungan akurasi sebagai berikut :

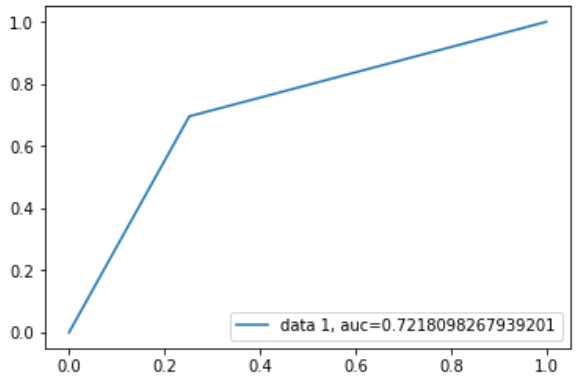
Auc = (TP+TN) / (TP + TN + FP + FN )) * 100%

= (48 +96) / (48 + 96 +27+21) * 100%

= 0.813

pada varibel di dalam grafik terdapat sebuah informasi data auc yang berarti ini adalah hasil akurasi metode Gradient Boosting memiliki nilai (0.813).

3.2. Decision Tree




*Gambar 10 curva roc_auc algoritma Decision Tree*

Gambar 10 adalah preview hasil akurasi metode Decision Tree. Grafik yang dijelaskan pada gambar 10 adalah grafik liniear dengn warna biru untuk varibel horizontal adalah range angka terkecil akurasi sampai terbesar dan untuk variabel vertikal sama berarti range angka terkecil akurasi terkecil sampai terbesar dengan perhitungan akurasi sebagai berikut :

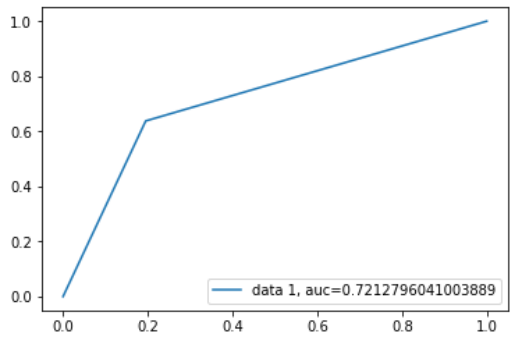
Auc = (TP+TN) / (TP + TN + FP + FN )) * 100%

= (48 +92) / (48 + 92 +32+21) * 100%

= 0.72

pada varibel di dalam grafik terdapat sebuah informasi data auc yang berarti ini adalah hasil akurasi metode Decision Tree memiliki nilai (0.72).

3.3.Random Forest




*Gambar 11 curv roc_auc algoritma Random Forest*

Gambar 11 adalah preview hasil akurasi metode Random Forest. Grafik yang dijelaskan pada gambar 11 adalah grafik liniear dengn warna biru untuk varibel horizontal adalah range angka terkecil akurasi sampai terbesar dan untuk variabel vertikal sama berarti range angka terkecil akurasi terkecil sampai terbesar dengan perhitungan akurasi sebagai berikut :

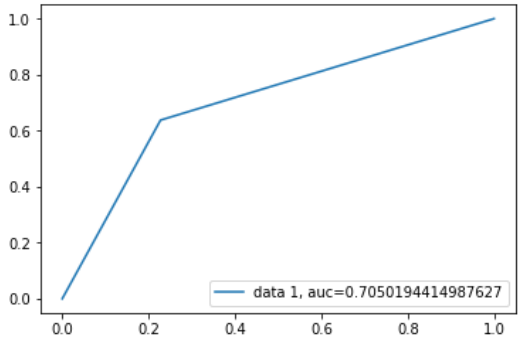
Auc = (TP+TN) / (TP + TN + FP + FN )) * 100%

= (44 +99) / (44 + 99 +24+25) * 100%

= 0.72

pada varibel di dalam grafik terdapat sebuah informasi data auc yang berarti ini adalah hasil akurasi metode Decision Tree memiliki nilai (0.72).

3.4.Logistic Regresion




*Gambar 12 curva roc_auc algoritma Logistic Regresion*

Gambar 12 adalah preview hasil akurasi metode Logistic Regresion. Grafik yang dijelaskan pada gambar 12 adalah grafik liniear dengn warna biru untuk varibel horizontal adalah range angka terkecil akurasi sampai terbesar dan untuk variabel vertikal sama berarti range angka terkecil akurasi terkecil sampai terbesar dengan perhitungan akurasi sebagai berikut :

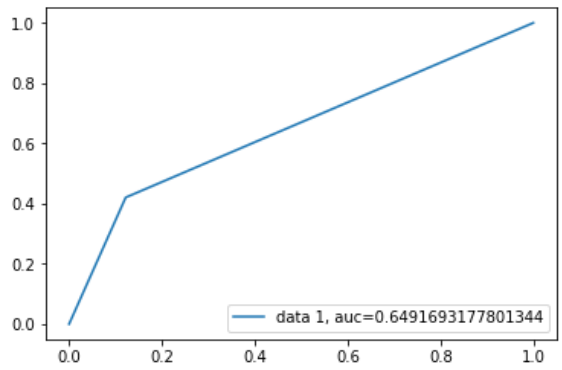
Auc = (TP+TN) / (TP + TN + FP + FN )) * 100%

= (44 +95) / (44 + 95 +28+25) * 100%

= 0.7

pada varibel di dalam grafik terdapat sebuah informasi data auc yang berarti ini adalah hasil akurasi metode Logistic Regresion memiliki nilai (0.7).

3.5. K-NN




*Gambar 13 curva roc_auc algritma K-NN*

Gambar 13 adalah preview hasil akurasi metode K-Nearest Neighbors. Grafik yang dijelaskan pada gambar 13 adalah grafik liniear dengn warna biru untuk varibel horizontal adalah range angka terkecil akurasi sampai terbesar dan untuk variabel vertikal sama berarti range angka terkecil akurasi terkecil sampai terbesar dengan perhitungan akurasi sebagai berikut :

Auc = (TP+TN) / (TP + TN + FP + FN )) * 100%

= (29 +108) / (29 +108+15+40) * 100%

= 0.64

pada varibel di dalam grafik terdapat sebuah informasi data auc yang berarti ini adalah hasil akurasi metode K-Nearest Neighbors memiliki nilai (0.64).

## 4. KESIMPULAN

Berdasarkan hasil uji coba teknik supervised learning dengan perbandingan 5 metode yaitu K-NN, Logistic Regresion, Random Forest , Decision Tree , Gradient Boosting dapat di lihat hasil akurasi dari 5 metode, bahwa metode yang paling akurat untuk prediksi diabetes dengan teknik supervised learning data pima-indian adalah metode Gradient Boosting dan untuk akurasi paling buruk akurasi adalah metode K-NN, dan untuk hasil prediksi metode decision tree, logistic regresion dan random forest hasil,nya hampir sama Untuk selanjutnya sebaiknya gunakan data yang lebih banyak untuk melatih model karena dalam machine learning semakin banyak data yang di gunakan dalam melatih model maka model akan semakin baik.

Dari hasil tersebut menghasilakan sebuah analisa mengapa metode gradient bosting lebih akurat karena ada beberapa hal yang mmepengaruhi sebagai berikut:

1. Pengaruh teknik cross validasi yang di tentukan data split dan data testing yg sama. Disini bisa di ketahui perbedaan ketika masuk dalam tahap cross validation hasil dari confusion matrik menunjukan angka dari hasil metode gardient bosting dg hasil TP"true positif" paling tinggi dari pada metode yang lain "48" meskipun metode Decision Tree juga menghasilkan angka sama pada hasil TP, akan tetapi di metode Decision Ttree FN " false negative" atau miss akurasi lebih banyak dari pada metode Gradient Boosting.
2. Di jelaskan dalam salah satu jurnal milik Jordan Frery, Amaury Habrard, Marc Sebban, Olivier Caelen, and

Liyun He-Guelton, yang berjudul tentang “Optimasi peringkat atas yang efisien dengan peningkatan gradien untuk deteksi anomali yang diawasi” di jurnal tersebut juga membuktikan bahwa untuk pencarian metode dengan teknik supervised learning paling efisien dan akurat adalah teknik boosting dan dari metode yang digunakan di atas untuk perbandingan akurasi teknik boosting ada pada metode Gradient Boosting [7].

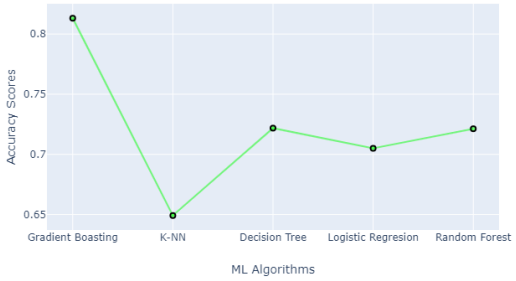
**3.** Artikel yang di tulis oleh Albolfazl Ravanshad “Data Scientist, Ph.D. dari university of florida dan beliau alah lulusan program nano degre machine learning Udacity. Menjelaskan tentang performa kinerja Gradient Boosting dengan metode Random Forest bahwa

a. Gradient Boosting: GBT membuat pohon satu per satu, di mana setiap pohon baru membantu memperbaiki kesalahan yang dilakukan oleh pohon yang sebelumnya dilatih
b. Kekuatan model : Karena pohon yang ditingkatkan diturunkan dengan mengoptimalkan fungsi objektif, pada dasarnya GBM dapat digunakan untuk menyelesaikan hampir semua fungsi objektif yang dapat di tulis gradien. Ini termasuk hal-hal seperti pemeringkatan dan regresi poisi, yang RF lebih sulit untuk dicapai.
c. Kelemahan model GBM lebih sensitif terhadap overfitting jika datanya berisik.

Pelatihan umumnya memakan waktu lebih lama karena fakta bahwa pohon dibangun secara berurutan.

GBM lebih sulit diatur daripada RF. Biasanya ada tiga parameter: jumlah pohon, kedalaman pohon dan tingkat pembelajaran, dan setiap pohon yang dibangun umumnya dangkal.

Random Forest : RF melatih setiap pohon secara mandiri, menggunakan sampel data acak. Keacakan ini membantu membuat model lebih kuat dari pada pohon keputusan tunggal, dan lebih kecil kemungkinannya untuk mengunakan data pelatihan [8].




*Gambar 14 hasil compasari 5 algotirm*

## 5. DAFTAR PUSTAKA

[1] B. S. D. Soumya, "J. Diabetes Metab.,," *Late stage complications of diabetes and insulin resistance,* vol. 2 (167), pp. pp. 2-7, 2011.

[2] V. K. Harleen Kaur, *Predictive modelling and analytics for diabetes using a machine learning,* Vols. Department of Computer Science and Engineering, School of

Engineering Sciences and Technology, Jamia Hamdard, New Delhi, India, 2018.12.004.

[3] U. s. K. C. o. T. S.Saru, "International Journal of Emerging Technology and Innovative Engineering," *ANALYSIS AND PREDICTION OF DIABETES USING MACHINE LEARNING,* vol. Volume 5, no. Issue 4, p. (ISSN: 2394 – 6598), April 2019.

[4] T. Y. e.-m. t. Kamer Kayaer e-mail: kayaer@yildiz.edu.tr, "Medical Diagnosis on Pima Indian Diabetes," *Medical Diagnosis on Pima Indian Diabetes Using General Regression Neural Networks,Yildiz Technical University , Department of Electronics and Comm. Eng. Besiktas, Istanbul 34349 TURKEY .*

[5] D. Chaturvedi, "Mathematical Models, Methods and Applications," *Soft computing techniques and their applications,* p. 31–40. Springer Singapore, 2015.

[6] M. S. A. G. K. Sajida Perveena, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes," *Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia ,* vol. Procedia Computer Science 82 ( 2016 ) 115 – 121 , pp. aDepartment of Computer Science & Engineering, University of Engineering & Technology, Lahore, Pakistan bTed Rogers School of Information Technology Management, Ryerson University, Toronto, Ontario, Canada cUniversity of Victoria, School of Health Informa, 30 March 2016.

[7] A. H. M. S. O. C. a. L. H.-G. Jordan Frery, "Efficient top rank optimization with gradient boosting," *Efficient top rank optimization with gradient boosting for supervised anomaly detection,* pp. 1 Univ. Lyon, Univ. St-Etienne F-42000, UMR CNRS 5516, Laboratoire Hubert-Curien, France .

[8] A. Ravanshad, "Gradient Boosting vs Random Forest," *https://medium.com/@aravanshad/gradient-boosting-versus-random-forest-cfa3fa8f0d80,* Apr 28, 2018.