PERBANDINGAN AKURASI ALGORITMA MACHINE LEARNING UNTUK PREDIKSI PENDAFTAR MAHASISWA BARU DI UNIVERSITAS XYZ

Supangat[1], Muhammad Rohmattullah Joyonegoro[2]

Program Studi Teknik Informatika, Fakultas Teknik, Universitas 17 Agustus 1945 Surabaya.
Jl. Semolowaru No. 45 Surabaya, 60118, Telp: (031) 5931800, Fax: (031) 5927817

Email: supangat@untag-sby.ac.id[1] , rohmat_bsi@untag-sby.ac.id[2]

Abstract

Admission of New school year students can be improved and can also be decreased. This is an issue facing the university in determining the future strategy steps. So there is a need for predictions or forecasting to determine the acquisition of new students, so that all policies and decisions in making future planning can be fulfilled properly.

This research aims to identify and predict the number of new student registries using Machine learning. The Dataset used was the number of new student registries in the year 2018-2020 XYZ University with a total of 7485 data that had been normalized with 6 feature Data and 1 Data Label. The algorithms used in this predisksi are Gradient Boosting, Decision Tree, K-NN, Logistic Regresion and Random Forest. Where the prediction results can provide ease to the university in determining the strategy measures in making decisions and policies in the coming year.

Keywords*: Machine Learning, Performa Accuration Method, Student Enrollment, Supervised Learning*

Abstrak

Penerimaan mahasiswa tahun ajaran baru dapat mengalami peningkatan dan dapat juga mengalami penurunan. Hal ini merupakan suatu masalah yang dihadapi Universitas dalam menentukan langkah-langkah strategi kedepannya. Sehingga diperlukan adanya prediksi atau peramalan untuk mengetahui perolehan jumlah mahasiswa baru, agar semua kebijakan dan keputusan dalam menyusun perencanaan kedepan dapat terpenuhi dengan baik.

Penelitian ini bertujuan untuk mengindentifikasi dan memprediksi jumlah pendaftar mahasiswa baru dengan menggunakan Machine learning. Dataset yang digunakan adalah jumlah pendaftar mahasiswa baru pada tahun 2018-2020 Universitas XYZ dengan jumlah 7485 data yang telah dinormalisasi dengan 6 Data Fitur dan 1 Label Data. Algoritma yang digunakan dalam predisksi ini adalah Gradient Boosting, Decision Tree, K-NN, Logistic Regresion dan Random Forest. Dimana hasil prediksi tersebut dapat memberikan kemudahan kepada Universitas dalam menentukan langkah-langkah strategi dalam mengambil keputusan dan kebijakan pada tahun yang akan datang.

Kata kunci*: Machine Learning, Performa Accuration Method, Student Enrollment, Supervised Learning*

1. PENDAHULUAN

Universitas merupakan sebuah tempat di mana berlangsungnya sebuah proses belajar-mengajar. Proses belajar-mengajar tersebut melibatkan peran serta sumber daya manusia yaitu Dosen dan Mahasiswa. Universitas memiliki sarana dan prasarana kuliah seperti ruang praktek dan ruang kuliah. Sarana dan prasarana merupakan hal penting dalam proses belajar-mengajar yang mampu mempengaruhi hasil akademik mahasiswa.

Machine Learning adalah Serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data dengan melakukan penggalian pola-pola dari data dengan tujuan untuk memanipulasi data menjadi informasi yang lebih berharga yang diperoleh dengan cara mengekstraksi dan mengenali pola yang penting atau menarik dari data yang terdapat dalam basis data.

Lembaga pendidikan formal setiap tahun rutin mengadakan kegiatan penerimaan mahasiswa baru. Jumlah mahasiswa pada penerimaan siswa tahun ajaran baru dapat mengalami peningkatan dan penurunan. Sehingga diperlukan adanya prediksi untuk mengetahui perolehan jumlah mahasiswa baru. Dengan harapan agar semua kebijakan dan keputusan dalam menyusun perencanaan akademik ke depan dapat terpenuhi dengan baik.

Machine Learning merupakan solusi berdasarkan permasalahan yang ada. Metode yang dilakukan adalah membandingkan setiap Algoritma agar bisa mengetahui prediksi. Dengan analisis perbandingan tersebut, diharapkan dapat membantu memprediksi jumlah mahasiswa yang mendaftar. Diharapan agar semua kebijakan dan keputusan dalam menyusun perencanaan akademik ke depan dapat terpenuhi dengan baik.

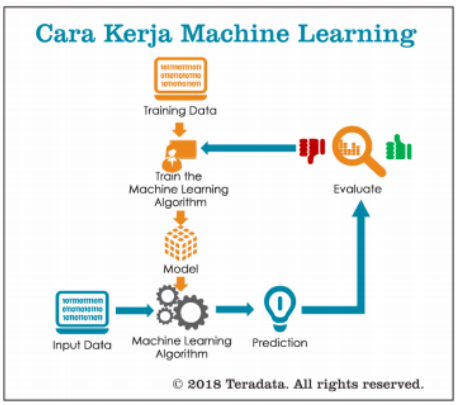
2. METODE PENELITIAN

A. Perancangan Sistem

Secara garis besar rancangan alus system machine learning sendiri terdiri dari beberapa tahap

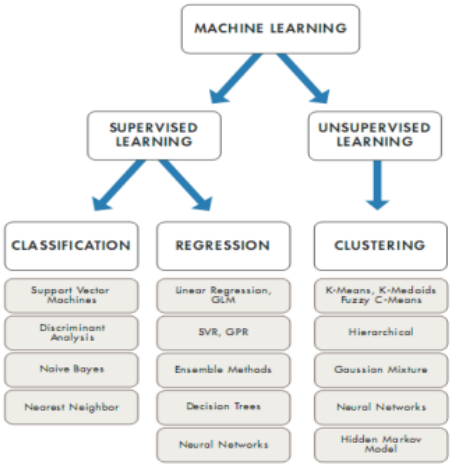
1. Mempersiapkan dataset
2. Normalisasi data
3. train test data
4. Cross validation
5. Ploting data.

Untuk alur system seperti gambar di bawah ini,



Gambar 1 Cara Kerja Machine Learning

B. Teknik Machine Learning

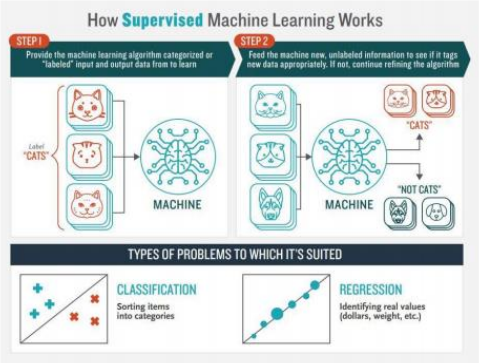


Gambar 2 Metode Pemisah Data Supervised dan Unsupervised

ploting Teknik machine learning dimana machine learning terbagi 2 teknik yaitu Teknik supervised learning dan unsupervised learning, disini karena saya pakai Teknik supervised learning metode

algoritma dan metode yang saya pakai berdasarkan algoritma classification dan regression untuk perbandingan akurasinya.

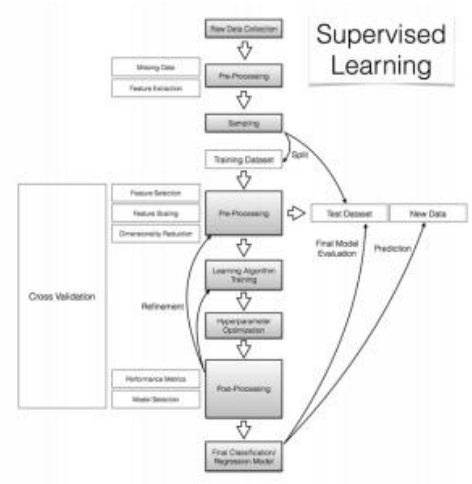
C. Cara Kerja Supervised Learning



Gambar 3 Cara Kerja Supervised Training

Tahap pertama supervised learning menyediakan data yang sudah di kategorikan atau "ada label" input dan output dari sebuah data untuk di gunakan dalam proses pembelajaran mesin. Tahap kedua machine learning akan melakukan pre-procesing untuk menentukan output apakah data tersebut merupakan jenis cats / no cats. Di dalam algoritma supervised learning terbagi menjadi 2 teknik yaitu klasifikasi dan regresi untuk mengembangkan model prediksi.

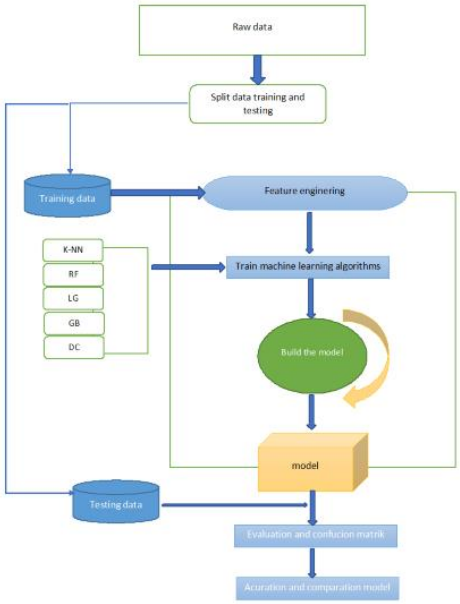
D. Flowchart Supervised Learning



Gambar 4 Flowchart Supervised Learning

Untuk flow alur supervised learning ini sendiri dimulai dari pertama siapkan data dan siapkan bentuk datanya berupa raw data setelah itu kemudian lakukan preprocesing dan dilanjutkan pengambilan data sampling untuk sampling disini dibagi data traning tadi untuk split atau bagi menjadi 2 data yaitu data training dan data testing, setalah itu lanjut ke tahap cros validation dimana di dalam cross validation terdapat tahapan prepocesing , learning algoritma traning, hyperparameter optimasi, post processing, dan setelah semua selesai lanjut ke tahap hasil final tentunya akurasi melalui confusion matrik.

E. Alur Sistem

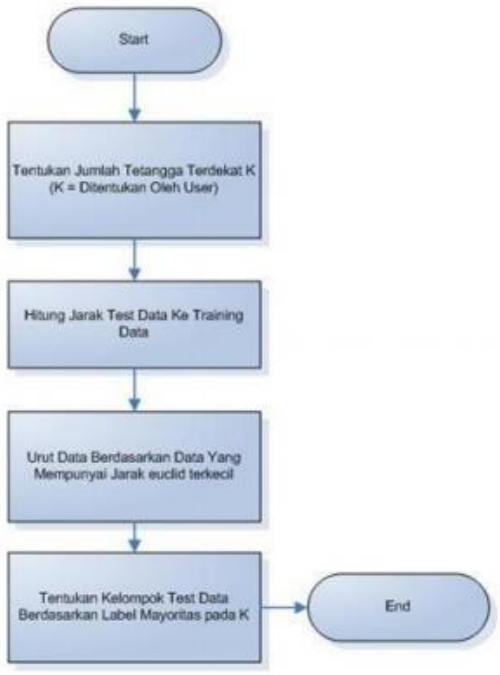


Gambar 5 Flowchart Alur Sistem

Untuk flow alur keseluruhan sistem ini sendiri dimulai dari pertama menyiapkan data dan data yang sudah di siapkan bentuknya adalah raw data setelah itu lakukan splitting dataset training dan testing, setelah mendapatkan data testing dan training lakukan tahap fitur engineering untuk memasukan fitur dari masing-masing algoritma dari data training dimana setiap algoritma memiliki fitur yang berbeda-beda, setelah itu masukan model dari tiap algoritma ke tahap train machine learning algoritma, dalam

tahap train machine learning ini masukan parameter nilai ambang batas dari semua fitur yang sudah di uji pada tahap sebelumnya, setelah mendapatkan parameter tersebut lakukan build model dari masing-masing algoritma, setelah tahap ini selesai tiap-tiap algoritma akan mendapatkan hasil output dari model tersebut, selanjutnya masukan data testing ke tahapan preprocesing dari model ke evaluation dan dapatkan hasil confusion matrik setelah itu dapatkan hasil acuration dari tiap metode dan comparasi akurasi dari 5 model yang telah di uji.

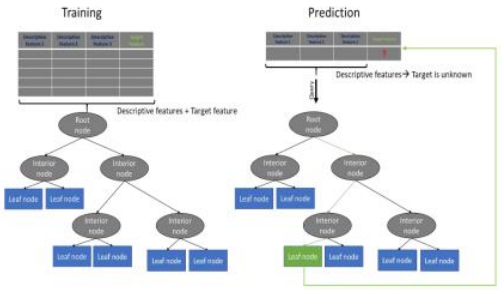
F. Algoritma K-NN



Gambar 6 Flowchart Metode K-NN

Untuk flow dari metode K-NN tahap pertama adalah menentukan jumlah K,nya, tahap selanjutnya lakukan perhitungan jarak data testing dan training, lanjut ke tahap selanjutnya yaitu mengurutkan data berdasarkan data yang mempunyai jarak euclid (perhitungan jarak dari 2 buah titik dalam Euclidean space. Euclidean space diperkenalkan oleh Euclid, seorang matematikawan dari Yunani sekitar tahun 300) terkecil, setelah itu tentukan kelompok data test berdasarkan label mayoritas pada K,

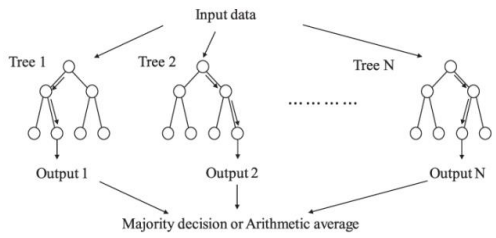
G. Algoritma Decision Tree



Gambar 7 Flowchart Metode Decision Tree

Flow model Decision Tree dibagi menjadi 2 ada flow model data training dan prediction, untuk flow model data training sendiri pertama lakukan import data, dalam bentuk raw data, setelah itu lakukan tahap root node dimana akan membagi sebuah data antara data training dan testing dan juga melakukan penyederhanaan data atau normalisasi untuk mengatasi jumlah data yang missing sampai menunjukan titik leaf node, untuk model yang ke dua flow prediction , flow prediction disini hampir sama dengan flow training hanya saja perbedaanya pada data karena di flow model prediction data label masih kosong maka untuk normalisasi data dengan output leaf node bisa di jadikan sebuah hasil data labelnya.

H. Algoritma Random Forest

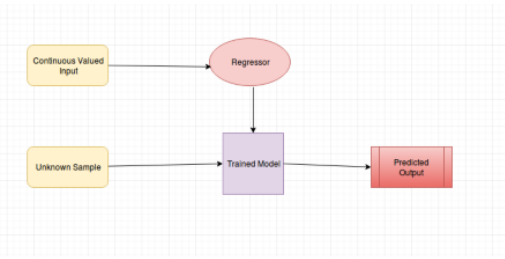


Gambar 8 Flowchart Metode Random Forest

Algoritma Random Forest adalah pengembangan dari Decision Tree jadi untuk alur hampir sama dimana proses normalisasi data akan di turunkan

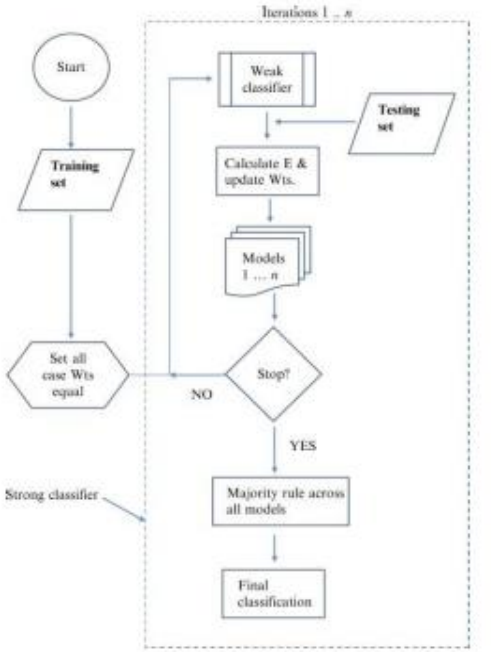
dengan hasil akhir beruapa leaf node sebagai outputnya akan tetapi di Random forest mempunyai perbedaan dimana sebuah tree yang di class,kan. Untuk tahap pertama input data setelah itu lanjutkan dengna tahap normalisasi berupa sebuah class sampai menemukan titik leaf node untuk di jadikan sebuah output.

I. Algoritma Linier Regresion



Gambar 9 Flowchart Metode Linier Regresion

J. Algoritma Gradient Boosting



Gambar 10 Flowchart Metode Gradient Boosting

Flow model Gradient Boosting di mulai dari menyiapkan data training, lanjut ke tahap pemilihan pengaturan setting case atau memasukan case uji coba, ketika proses memsukan case ini di lakukan maka selanjutnya tentukan aturan pada semua rule modenya setelah proses ini selesai otomatis mendapatkan result dari hasil atau final clasification, jika proses pemasukan case tidak dilakukan maka masuk ke tahap perhitungan clasfikasi terlebih dahulu dengan memasukan data testing, setelah itu lakukan update data dan lakukan proses data tersebut dengan iterasi model I-N sampai menghasilkan sebuah output klasifkasinya.

3. HASIL DAN PEMBAHASAN

Sebelum menjadi dataset, Data Penerimaan Mahasiswa baru memiliki banyak kekurangan. Yang pertama yaitu banyak data yang tidak sesuai dan data kosong sehingga mempengaruhi prediksi

3.1. Dataset

	kodependapatanayah	kodependapatanibu	penghasilanwali	sistemkuliah	kodekota	listsumber	pendaftar
0	1	1	3	2	4	3	1
1	3	1	3	1	4	1	0
2	4	4	1	1	3	3	1
3	1	1	1	1	4	1	1
4	2	2	2	2	1	3	0

Gambar 11 Dataset yang digunakan

Dataset penerimaan mahasiswa baru dengan 6 data fitur yang terdiri dari “kodependapatanayah, kodependapatanibu, penghasilanwali, sistemkuliah, kodekota, listsumber” dan 1 data lebl terdiri dari “pendaftar”. Pada gambar diatas menjelaskan tentang bentuk data berupa raw data, untuk dataset sendiri berjumlah 7484 raw.

3.2. Hasil Ujicoba metode

Tabel 1 Hasil Ujicoba Metode

No	Metode	Hasil Akurasi Data Training	Hasil Akurasi Data Testing
1	Gradient Boosting	0.925	0.922

2	Decision-Tree	0.926	0.922
3	K-NN	0.93	0.92
4	Logistic Regresion	0.925	0.922
5	Random Forest	0.925	0.922

. **3.3. Hasil Ujicoba data testing**

Tabel 2 Ujicoba 100 Data

No	Metode	Hasil Akurasi Metode
1	Gradient Boosting	0.31522
2	Decision Tree	0.41304
3	K-NN	0.45652
4	Logistic Regression	0.47826
5	Random Forest	0.47826

Pada Tabel 2 menjelaskan tentang hasil dari akurasi dengan menggunakan 100 data. Dengan hasil akurasi tertinggi yaitu Random Forest dan Logistic Regression dengan hasil (0.47826) dan akurasi yang paling rendah adalah Gradient Boosting dengan hasil (0.31522).

Tabel 3 Ujicoba 500 Data

No	Metode	Hasil Akurasi Metode
1	Gradient Boosting	0.47957
2	Decision Tree	0.58261
3	K-NN	0.47391
4	Logistic Regression	0.5
5	Random Forest	0.5

Pada Tabel 3 menjelaskan tentang hasil dari akurasi dengan menggunakan 500 data. Dengan hasil akurasi tertinggi yaitu Random Forest dan Logistic Regression dengan hasil (0.58261) dan akurasi yang paling rendah adalah Gradient Boosting dengan hasil (0.47957).

Tabel 4 Ujicoba 1500 Data

No	Metode	Hasil Akurasi Metode
1	Gradient Boosting	0.58421
2	Decision Tree	0.49931
3	K-NN	0.49564
4	Logistic Regression	0.5
5	Random Forest	0.5

Pada Tabel 4 menjelaskan tentang hasil dari akurasi dengan menggunakan 1500 data. Dengan hasil akurasi tertinggi yaitu Gradient Boosting dengan hasil (0.58421) dan akurasi yang paling rendah adalah K-NN dengan hasil (0.49564).

Tabel 5 Ujicoba 3000 Data

No	Metode	Hasil Akurasi Metode
1	Gradient Boosting	0.59219
2	Decision Tree	0.49682
3	K-NN	0.50787
4	Logistic Regression	0.5
5	Random Forest	0.5

Pada Tabel 5 menjelaskan tentang hasil dari akurasi dengan menggunakan 3000 data. Dengan hasil akurasi tertinggi yaitu Gradient Boosting dengan hasil (0.59219) dan akurasi yang paling rendah adalah K-NN dengan hasil (0.49682).

Tabel 6 Ujicoba 5000 Data

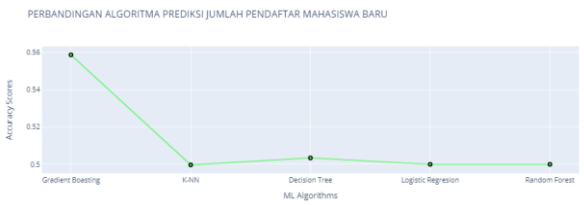
No	Metode	Hasil Akurasi Metode
1	Gradient Boosting	0.53303
2	Decision Tree	0.5
3	K-NN	0.49695
4	Logistic Regression	0.5
5	Random Forest	0.5

Pada Tabel 6 menjelaskan tentang hasil dari akurasi dengan menggunakan 5000 data. Dengan hasil akurasi tertinggi yaitu Gradient Boosting dengan hasil (0.53303) dan akurasi yang paling rendah adalah K-NN dengan hasil (0.49695).

Tabel 7 Ujicoba Seluruh Dataset

No	Metode	Hasil Akurasi Metode
1	Gradient Boosting	0.55865
2	Decision Tree	0.50627
3	K-NN	0.49971
4	Logistic Regression	0.5
5	Random Forest	0.5

Tabel 7 menjelaskan tentang hasil dari akurasi data yang sudah tervalidasi dari hasil tersebut bisa di pastikan untuk uji coba kali ini metode yang terbaik adalah Gradient Boosting dengan nilai akurasi (0. 55865). disusul dengan metode Decision Tree dengan hasil (0. 50627), urutan ketiga metode Random Forest dengan hasil (0.5), selanjutnya metode Logistic Regresion dengan hasil (0.5), untuk metode paling rendah akurasinya adalah K-NN dengan hasil (0. 49971).



Gambar 12 Grafik Perbandingan Algoritma

Penjelasan gambar 12 adalah kesimpulan dari ke 5 metode 1.Gradien Boosting, 2. Decision Tree, 3

K-NN, 4.Logistic Regresion, 5.Random Forest yang di visualisaikan berupa grafik liniear, dengan varibel horizontal memiliki sebuah keterangan dari kelima metode tersebut dan untuk variabel vertikal terdapat parameter accuray score yang berarti sebuah nilai angka dari metode dari yang paling rendah sampai nilai tertinggi.

4. SIMPULAN

Hasil akhir uji coba sudah di lakukan seperti tabel diatas, dapat disimpulkan :

Tingkat akurasi yang bagus adalah metode Gradient Boosting dan untuk metode dengan akurasi paling buruk adalah K-NN, disusul dengan hasil metode Logistic Regresion, Random Forest dan Decision Tree. Dari hasil tersebut menghasilakan sebuah analisa mengapa metode gradient bosting lebih akurat karena ada beberapa hal yang memepengaruhi 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.
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. Hutan Acak: 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.

DAFTAR PUSTAKA

BigsmiLe, M. (2016, Mei 17). *Mengenal Teknologi Machine Learning (Pembelajaran Mesin)*. Retrieved from Code Politan: https://www.codepolitan.com/mengenal-teknologi-machine-learning-pembelajaran-mesin

Bradley, J. (2015, Januari 21). *Random Forests and Boosting in MLlib*. Retrieved from Databricks: https://databricks.com/blog/2015/01/21/random-forests-and-boosting-in-mllib.html

Brownlee, J. (2019, April 16). *A Gentle Introduction to Scikit-Learn: A Python Machine Learning Library*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/a-gentle-introduction-to-scikit-learn-a-python-machine-learning-library/

Fajar, R. (2016, Juli 24). *Memulai Pemrograman dengan Python*. Retrieved from Code Politan: https://www.codepolitan.com/memulai-pemrograman-python

Jose, I. (2018, November 8). *KNN (K-Nearest Neighbors) #1*. Retrieved from Towards Data Science: https://towardsdatascience.com/knn-k-nearest-neighbors-1-a4707b24bd1d

Klein, B. (2018, Mei 18). *Boosting*. Retrieved from Python Machine Learning Tutorial: https://www.python-course.eu/Boosting.php

Klein, B. (2018, June 10). *What are Decision Trees*. Retrieved from Python Machine Learning Tutorial: https://www.python-course.eu/Decision_Trees.php

Narkhede, S. (2018, May 17). *Understanding Logistic Regression*. Retrieved from Towards Data Science: https://towardsdatascience.com/understanding-logistic-regression-9b02c2aec102