Jurnal Sistem dan Teknologi Informasi (JSTI)

https://journalversa.com/s/index.php/jsti Vol. 06, No. 3 Agustus 2024

PERBANDINGAN PENGGUNAAN STATISTIKA DAN MACHINE LEARNING DALAM TEKNIK PENGOLAHAN DATA

Muhammad Haris Husni[1]
Email: harishusni@student.unp.ac.id

Resman Hendi Nofanolo Harefa[2]
Email: resmanh88@student.unp.ac.id

Ambiyar[3]
Email: ambiyar@ft.unp.ac.id

Mahesi Agni Zaus[4]
Email: mahesiagnizaus@ft.unp.ac.id

[1,2,3,4]Universitas Negeri Padang

ABSTRAK

Saat ini kemajuan teknologi berkembang begitu cepat sehingga dunia bisa dikatakan sedang berada pada era transformasi digital, dari teknologi berskala besar hingga teknologi sederhana, yang tidak dapat dipikirkan oleh manusia. Kita dapat mengakses informasi dan data yang ingin kita ketahui seperti industri, politik, hiburan, dan lain-lain melalui Internet. Statistika berperan penting yaitu sebagai sarana bagi perkembangan ilmu lainnya, baik tentang alam maupun sosial. Statistika digunakan dalam bidang ilmu lainnya sebagai sarana untuk menarik kesimpulan, mengetest hipotesis, ide, mengetahui keadaan, melakukan analisis eksperimen, mengambil keputusan, dan lain-lain. Machine learning telah mendapatkan perhatian di bidang teknologi dalam beberapa tahun terakhir. Sistem ini diharapkan dapat mengubah dan meringankan life style dan pekerjaan orang. Machine learning dan statistik yaitu dua bidang penelitian ilmu data yang terkait, dan keduanya memiliki landasan ide dan model yang serupa. Perbedaan keduanya berada pada fokusnya yang berbeda. Walaupun statistik berfokus pada penarikan kesimpulan dan Machine learning berfokus untuk memperkirakan data baru dengan taraf akurat yang tinggi. Karena persamaan dan perbedaan itu, tidak mengapa jika kita menyebut statistika dan machine learning adalah dua hal yang berbeda dalam bidang keilmuan yang sama.

Kata Kunci: Machine Learning, Pengolahan Data, Statistika.

ABSTRACT

Currently, technological progress is developing so fast that the world can be said to be in an era of digital transformation, from large-scale technology to simple technology, which humans cannot think about. We can access information and data that we want to know, such as industry, politics, entertainment, etc. via the Internet. Statistics plays an important role, namely as a means for the development of other sciences, both natural and social. Statistics is used in other

fields of science as a means to draw conclusions, test hypotheses, ideas, determine situations, carry out experimental analysis, make decisions, and so on. Machine learning has gained attention in the technology field in recent years. This system is expected to be able to change and lighten people's life styles and work. Machine learning and statistics are two related areas of data science research, and both have similar underlying ideas and models. The difference between the two lies in their different focuses. Although statistics focuses on drawing conclusions and machine learning focuses on predicting new data with a high degree of accuracy. Because of these similarities and differences, it is okay if we say that statistics and machine learning are two different things in the same scientific field.

Keywords: *Machine Learning, Data Processing, Statistics.*

1. PENDAHULUAN

Kebanyakan oknum yang bekerja dibidang pengetahuan memiliki pengalaman di dalam bidang statistika. Tidak dapat dibayangkan ternyata statistika berperan penting sebagai sarana bagi kemajuan ilmu-ilmu lain, baik pengetahuan maupun ilmu-ilmu sosial. Statistika juga digunakan dalam bidang ilmu lainnya sebagai sarana untuk menarik kesimpulan, mengetest hipotesis, ide, memahami gejala, menelaah eksperimen, dan mengambil keputusan, dan sebagainya.

Machine learning masa ini merupakan salah satu bagian ilmu yang banyak dibahas di media. Karena merupakan merupakan bagian kecerdasan buatan (AI), nyaris semua orang telah menggunakannya atau bahkan mendengar tentang sistem komputer yang dibuat menggunakan teknik machine learning. Seperti anda dapat Lihat tag foto otomatis di Facebook, dapat mencari rekomendasi di Google, memilih rekomendasi produk serupa saat berbelanja online, gunakan layanan email bebas spam, dan banyak lagi.

Landasan dasar statistika dan machine learning adalah ilmu teori probabilitas. Seluruh metode dalam statistik dan pembelajaran mesin didasarkan pada teori probabilitas, sebuah bahasa matematika untuk menguji derajat ketidakpastian. Aljabar linier, analisis, dan teknik optimasi merupakan keahlian dasar yang penting untuk statistika dan machine learning. Pengetahuan mendasar ini memberikan landasan untuk pengumpulan data, pemrosesan data, dan teknik analisis data, serta untuk memahami tantangan yang ditimbulkan oleh pentingnya pembelajaran mesin di masa depan.

2. LANDASAN TEORI

Probaliitas

Ada banyak ketidakpastian di dunia ini. Faktanya, machine learning berhubungan dengan ketidakpastian. Dengan cara ini, pembelajaran mesin sangat kuat kaitannya dengan statistika. Probabilitas memberikan kerangka kerja untuk mengukur dan memanipulasi ketidakpastian. Menurut definisinya, probabilitas adalah kemungkinan terjadinya suatu peristiwa dengan memprediksi peristiwa mana yang akan terjadi. Oleh karena itu, nilai probabilitas hanya dapat berkisar antara 0 hingga 1. Nilai 0 berarti keadaan tersebut tidak terjadi, meskipun nilai 1 menunjukkan keadaan tersebut terjadi. Probabilitas bisa digunakan untuk memprediksi keadaan yang akan berlangsung dan menetapkan keputusan yang akan dipilih (Murphy, 2012).

Statistika

Statistika adalah ilmu dasar machine learning dan memungkinkan anda memahami fenomena, membuat keputusan, dan mengetest hipotesis berdasarkan data. Dan statistika dapat memecahkan masalah atau menetapkan hasilnya dengan mencari pola yang nilai kesalahannya paling kecil agar hasil yang diperoleh benar. Statistika adalah cabang matematika yang secara spesifik membahas bagaimana data disatukan, dianalisis, dan dirumuskan. Dengan kata lain, istilah “statistika” di sini mengacu pada kumpulan ilmu pengetahuan tentang metode pengambilan sampel, pengumpulan data, serta analisis dan interpretasi data (Furgon, 2009).

Statistika hanya berfungsi sebagai bantuan, peran statistika dalam penelitian adalah sebagai alat, artinya tujuan statistika bukanlah acuan untuk menentukan komponen penelitian lainnya. Oleh karena itu, penting untuk menentukan pertanyaan yang ingin Anda jawab dan tujuan penelitian tersebut. Statistika membantu dalam membangun model, merumuskan hipotesis, mengembangkan alat pengumpulan data, menyiapkan rencana penelitian, menentukan sampel, menganalisis data, dan menafsirkan data dengan cara yang bermakna.

Nyaris semua penelitian ilmiah djalankan berdasarkan sampel peristiwa, dan generalisasi dibuat berdasarkan sampel tersebut. Kesalahan tidak bisa dihindari ketika melakukan generalisasi sehingga di sinilah beberapa fungsi statistik bekerja berdasarkan sampel, tidak populasi. Oleh karena itu hipotesis dapat diuji menggunakan metode statistik. Hasil analisis statistik yang diperoleh berdasarkan perhitungan numerik tersebut tidak ada artinya kecuali jika benar-benar dituliskan dalam bentuk kalimat atau kata dan ditarik suatu kesimpulan. Jika tidak, maka hasil analisisnya tidak ada artinya, hanya menyisakan angka “tidak berdering” (Suggyono, 2014).

3. HASIL DAN PEMBAHASAN

Kemajuan statistik dan machine learning pastinya tidak lepas dari elemen penting yaitu data. Machine learning sekurang-kurangnya memiliki dua maksud utama yaitu tentang memecahkan masalah dan/atau memperoleh pengetahuan dalam memperkirakan masa depan (peristiwa yang tidak teramati). Machine learning statistik mengarah pada teknologi yang memprediksi masa depan dan secara rasional mengekstraksi pengetahuan dari data. machine learning statistik mungkin merupakan alat atau metode yang cocok untuk mencapai tujuan ini. Meskipun statistik berfungsi sebagai landasan pembelajaran yang menerepkan teori statistik untuk menyimpulkan dan mengartikan model, sedangkan machine learning berfokus pada pemanfaatan model untuk memprediksi data baru.

Materi yang dibahas pada statistika dan machine learning hampir sama, dan pengetahuan pokok yang diperlukan juga sama. Lalu apa perbedaan antara statistik dan machine learning, perbedaan pentingnya terletak pada perbedaan orientasi kedua bidang tersebut. Statistik berfokus pada penarikan kesimpulan dan interpretasi model, sedangkan pembelajaran mesin berfokus pada pemanfaatan model untuk mengetahui data baru. Pada asal mulanya, pembelajaran mesin mengizinkan komputer mempelajari dan membedakan alur tanpa memerlukan pemrograman eksplisit. Kombinasi metode statistik dan machine learning dapat digunakan sebagai alat yang ampuh untuk mengetahui berbagai jenis data di banyak bidang ilmu komputer, seperti pemrosesan gambar, pemrosesan bahasa, pemrosesan bahasa natural, dan kontrol robot, serta dalam bidang ilmu dasar, seperti biologi. Kedokteran, dan sumber daya.

Data merupakan salah satu elemen kunci dalam pemodelan, jadi sebelum kita melanjutkan ke langkah pemrosesan data, kita perlu mengetahui dahulu jenis-jenis data. Dalam statistical machine learning, pemilihan sampel data sangatlah penting. Jika data sample tidak dapat mewakili populasi, maka model pembelajaran yang dihasilkan tidak sesuai. Karena alasan ini, data pengujian biasanya juga dibuat. Mesin diarahkan untuk menggunakan data sample, lalu kinerjanya diuji menggunakan data uji. Keterwakilan suatu populasi dapat diketahui dengan memeriksa ciri-ciri data tersebut. Ringkasnya, istilah “training” adalah proses membuat suatu model, dan istilah “pengujian” adalah tahap pengujian kriteria model yang dipelajari. Dataset yaitu kumpulan data (sampel dalam statistik). Contohnya yaitu data yang digunakan untuk membuat model dan menilai model machine learning.

Mengola Data menggunakan Machine Learning

Ada dua sebutan penting saat membuat model machine learning, training adalah proses membangun model, dan pengujian adalah proses mengetest kinerja model yang learning.

Dataset adalah gabungan data (sampel dalam statistik). Contoh ini adalah data yang dimanfaatkan untuk membuat model dan menilai model pembelajaran mesin. Secara umum, dataset dikategorikan menjadi tiga tipe yang tidak beririsan (sampel dalam kumpulan tertentu tidak muncul di kumpulan lain).

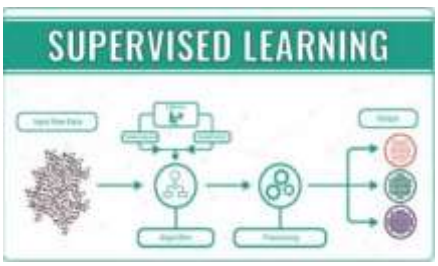
A. Training set adalah kelompok data untuk melatih atau membangun model.
B. Development set adalah kelompok data untuk pengoptimalan saat melatih model yaitu training set, dan performanya kemudian diuji selama pelatihan pada set validasi. Hal ini membantu generalisasi sehingga model dapat mengenali pola secara umum.
C. Testing set adalah kelompok data untuk mengetest model setelah proses pelatihan selesai.

Sampel dalam kumpulan data disebut titik data dan mewakili peristiwa statistik. Proses training, validasi, dan uji idealnya berasal dari distribusi (sampel) yang serupa dan memiliki keistimewaan yang serupa (independent and identically distributed). Ada tiga jenis Machine Learning yaitu supervised learning, unsupervised learning, & reinforcement learning.

a. supervised learning

Dalam algoritma tersebut, kumpulan data training diberikan kepada sistem dalam bentuk informasi masukan dan keluaran yang diperlukan agar sistem dapat belajar melalui data yang ada. Ketika sistem mencari suatu pola dalam suatu kumpulan data, pola tersebut digunakan sebagai acuan untuk data selanjutnya. Cara kerja pembelajaran yang diawasi: Model pembelajaran yang diawasi adalah model yang digunakan untuk memperkirakan hasil di masa akan datang memalui data historis yang ada.

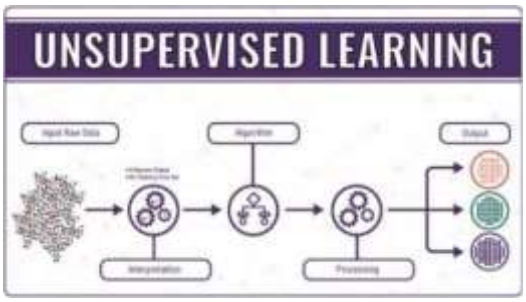
Dalam model ini, pertama-tama Anda diberikan petunjuk tentang cara mempelajari sesuatu. Misalnya, model supervised learning menggunakan algoritma yang ada untuk memprediksi kemungkinan terjadinya bencana alam seperti gempa bumi dan tsunami. Model ini juga dapat diartikan sebagai pendekatan data terlatih. Selain itu, supervisi learning juga menandai variabel untuk mengelompokkan data menjadi informasi yang ada.



Picture 1. Supervised Learning

b. Unsupervised learning

Berbeda dengan supervised, algoritma ini tidak bersifat prediktif dan tidak menerima training data set karena harus belajar dari data yang ada. Namun algoritma tersebut bersifat deskriptif dan cocok untuk menyusun atau mengklasifikasikan data. Seperti itu cara kerja algoritma unsupervised learning. Dalam model unsupervised learning tidak ada tujuan atau variabel yang tetap. Selain itu, praktisi data tidak memerlukan keahlian khusus untuk prediksi terhadap data. Selain itu, model unsupervised learning memungkinkan algoritma menemukan pola yang sembunyi di dalam data itu sendiri. Contoh model ini adalah menciptakan segmentasi pasar untuk menjalankan campaign yang efektif berdasarkan pengelompokan.



Picture 2. Unsupervised Learning

c. Reinforcement Learning

Ide dasar dari algoritma ini adalah terdapat agen ditempatkan di lingkungan yang tidak diketahui dan mengeksplorasi seluruh lingkungan untuk menemukan konsep reward dan error. Jika agen menemukan kesalahan, ia akan terus mencari cara hingga mendapatkan imbalannya. Pencarian ini dilakukan melalui trial and error. Agen terus bereksperimen tetapi tidak melakukan kesalahan apa pun pada soal yang sama. Dalam model pembelajaran penguatan, mesin dilatih untuk membuat keputusan spesifik berdasarkan kebutuhan bisnis dengan tujuan memaksimalkan kinerja. Model ini memastikan bahwa perangkat lunak atau mesin terus dilatih sesuai dengan daerah yang diakibatkannya. Tidak hanya itu, model ini juga digunakan untuk menyelesaikan permasalahan dalam perusahaan.

Picture 3. Reinforcement learning

Ada banyak teknik dan langkah berbeda untuk mengimplementasikan machine learning,

namun secara umum proyek machine learning terdiri dari tahapan berikut:

A. Mendefinisakan Masalah

Hambatan terjadi ketika target yang diinginkan tidak berhasil (keadaan saat ini bukanlah keadaan yang diinginkan). Untuk membawa situasi saat ini ke keadaan yang diinginkan, kita harus melakukan aktivitas yang disebut pemecahan masalah. Setiap field (domain) mempunyai definisi masalah yang berbeda-beda. Sebab itu, mengetahui Teknik machine learning tidak mengetahui domain aplikasinya bukanlah hal yang baik.

B. Mengumpulkan Data

Ini dibentuk dengan menyimpan banyak file yang berisi data yang dapat diperiksa oleh komputer dan dibagi menjadi fungsi keluaran dan masukan.

C. Menyiapkan Data

Menentukan kualitas data pada preprocessing data agar hasilnya optimal/baik.

D. Menjelaskan data sebagai feature vector

Tahap ini dilakukan dengan memilih algoritma yang sesuai dan merepresentasikan data berupa bentuk model.

E. Menguji model dan mengetahui kinerjanya menggunakan data validasi (development).

F. Melakukan pengujian dan analisis model kuantitatif dan kualitatif. Tahap ini mengetest keakuratan hasil terhadap bagian kumpulan test dataset.

G. Presentasi hasil.

Learning Pengolahan Data pada Statistika

Dalam peraturan Nomor 16 Tahun 1997 mengenai Statistika, arti statistik mempunyai tiga aspek yaitu lmu atau metode mempelajari data atau informasi dalam bentuk angka, sistem yang menggabungkan penerapan statistik, dan metode pengumpulan, pengolahan, menamapilkan dan menganalisis data. Sedangkan Statistik adalah ilmu menyusun data dan probabilitas. Statistik mengajarkan kita untuk menyusun data dan menghitung prediski yang akan terjadi.

Target utama penelitian adalah untuk memperoleh jawaban pertanyaan penelitian. Untuk mencapai tujuan tersebut, antara lain diperlukan pengolahan dan menelaah data.

Langkah-langkah mengolah data statistik adalah sebagai berikut:

a. Pengumpulan Data

Sebelum pengolahan data, beberapa langkah yang harus dibuat. Sedangkan sesudah dilakukan analisis data yaitu tahap pendekatan data, sehingga data tersebut dengan mudah diinterpretasikan. Kuesioner merupakan salah satu alat pengumpulan data untuk survei. Sebaiknya disediakan kolom koding pada kuesioner untuk memudahkan proses selanjutnya.

b. Editing Data

Anda perlu mengedit data lapangan yang ada, seperti survei. Tujuannya adalah untuk memeriksa apakah survei telah selesai, apakah jawabannya logis, dan apakah terdapat konsistensi antar pertanyaan.

c. Koding Data

Tahap ini berlaku untuk soal tertutup (pengkodean dilakukan sebelum masuk kepengolahan), soal semi terbuka dan soal terbuka.

d. Pengolahan Data

Pemrosesan data melibatkan dua tugas. Artinya, memasukkan data pada saat proses entri atau agregasi data dan mengedit kembali data gabungan untuk menghindari kesalahan entri data atau penempatan kolom atau baris yang salah dalam tabel.

e. Analisis dan interpretasi data

Saat melakukan analisis data, penting untuk mengetahui cara menggunakan alat analisis dengan benar. Hal ini karena meskipun alat analisisnya sangat canggih, namun jika prinsip penggunaannya tidak diikuti, hasil yang diperoleh dapat disalahartikan. Tidak ada gunanya menarik kesimpulan. Model statistik untuk tujuan analisis data telah berkembang dari model statistik deskriptif menjadi statistik inferensial nonparametrik. Saat memutuskan untuk menggunakan alat statistik untuk melakukan analisis data, ada beberapa hal yang perlu dipertimbangkan:

a. Bagaimana datanya didapat, apakah didapat dari sampel (menggunakan metode sampling) atau dari populasi (menggunakan sensus)?
b. Jika dari sampel, apa metode sampling yang digunakan, kelompok sampel probabilitas atau kelompok sampel non-probabilitas? Pada nilai apa data diukur (nominal, ordinal, interval, rasio, dll)?

c. Bagaimana merumuskan hipotesis, apakah sebaiknya melakukan uji satu sisi atau dua sisi bila menggunakan statistik inferensial.

Perbedaan Teknik pengolahan data dengan statistika dan machine learning

Gambaran sederhana statistical machine learning dapat diilustrasikan dengan variabel terikat kategoris y yang dianalisis menggunakan beberapa variabel bebas x1,... . . xm. Contoh penerapannya adalah analisis surat keterangan sekolah sekolah dimana y = 1 berarti 'lulus' dan y = 0 berarti 'gagal'. Variabel Xi meliputi profil siswa, total waktu belajar, nilai mata pelajaran, aktivitas di kelas, dan variabel lain yang berkaitan dengan kelulusan. Teknik yang umum digunakan untuk memodelkan masalah tersebut adalah dengan menggunakan model regresi logistik.

Statistika dan machine learning membuat sebuah konsep yang disebut Statistical machine learning dengan menerapkan model regresi logistik. Meskipun Rata-rata peneliti statistik memiliki latar belakang matematika, sedangkan peneliti machine learning memiliki latar belakang algoritma. Namun, statistik dan machine learning menerapkan dasar dan rumus yang sama untuk menyelesaikan masalah. Keduanya membahas konsep variabel acak, distribusi statistik, expected value, variansi, sampai pada konsep distribusi prior dan posterior.

Statistika dan machine learning membentuk sebuah konsep yang disebut Statistical machine learning dengan menggunakan model regresi logistik. Meskipun Rata-rata peneliti statistik memiliki latar belakang matematika, sedangkan peneliti machine learning memiliki latar belakang algoritma. Namun, statistik dan machine learning menggunakan teori dan rumus yang sama untuk menyelesaikan masalah. Keduanya membahas konsep variabel acak, distribusi statistik, expected value, variansi, sampai pada konsep distribusi prior dan posterior.

Alur kerja ahli statistik saat membangun model regresi logistik biasanya memeriksa asumsi, menggunakan perangkat lunak untuk melakukan estimasi parameter (maximum likelihood), memeriksa nilai parameter dan signifikansinya, dan mengidentifikasi variabel yang signifikan. Untuk melihat dan membandingkan model mana yang lebih baik (variabel mana yang sebaiknya dimasukkan ke dalam model), digunakan ukuran seperti Akaike Information Criterion (AIC). Setelah model dibuat, ahli statistik menganalisis interpretasi parameter yang dihasilkan, seperti: Variabel apa saja yang mempengaruhi kelulusan, apakah berdampak positif atau negatif, dan seberapa besar dampaknya (dipastikan dengan menggunakan nilai parameter variabel tersebut).). Pada akhirnya, para ahli statistik menarik kesimpulan tentang hubungan antara variabel-variabel ini dan kelulusan siswa.

Contoh masalah serupa yang dihadapi oleh praktisi machine learning, dengan variabel terikat kategoris y dan beberapa variabel bebas x1. . . . :xm. Contoh penerapannya adalah deteksi email spam. Nilai variabel y = 1 berarti email spam, y = 0 berarti tidak ada email spam. Variabel xj berisi karakteristik email tersebut. Misalnya, jumlah kata dalam email, apakah email tersebut berisi lampiran, apakah email tersebut memiliki gambar, atau apakah email tersebut berisi kata-kata tertentu (seperti 'promosi' atau "obat'. Model regresi logistik juga merupakan salah satu model yang umum digunakan oleh para ahli machine learning . Mereka juga menggunakan perangkat lunak untuk memperkirakan parameter maximum likelihood.

Dari kedua alur kerja di atas, terlihat jelas bahwa meskipun sama-sama menggunakan model regresi logistik, namun terdapat perbedaan fokus ahli statistik dan machine learning. Selain perbedaan fokus (interpretasi vs prediksi), terdapat beberapa perbedaan penting. Ini termasuk cara praktisi machine learning menangani estimasi parameter, tidak seperti ahli statistik yang kurang peduli dengan nilai parameter tersebut. praktisi machine learning lebih memilih untuk menjaga nilai parameter ini relatif kecil dengan menambahkan regularisasi. Namun, hal ini tidak menjadi masalah selama model dapat memprediksi data baru secara akurat, berapa pun nilainya. Perbedaan lainnya adalah cara pemilihan model. Pakar statistik cenderung memilih model berdasarkan teori analitik seperti AIC, sedangkan pakar pembelajaran mesin cenderung memilih model berdasarkan performa empirisnya dalam fase validasi silang.

4. KESIMPULAN

Ada banyak persamaan dan perbedaan antara bidang statistik dan machine learning. Kedua bidang tersebut didasarkan pada teori probabilitas dan membahas landasan teori dan model yang sama. Perbedaan keduanya terletak pada fokusnya yang berbeda. Meskipun statistik berfokus pada penarikan kesimpulan, machine learning berfokus pada memprediksi data baru. machine learning dengan kemampuan mencari pola dalam data. Jika pola-pola pada data diketahui, maka machine learning dapat menarik kesimpulan berupa pola-pola yang terbentuk, sehingga keputusan dapat diambil setelah kesimpulan tersebut ditentukan. machine learning memiliki sifat prediktif dan deskriptif yang membuktikan kemampuannya. Ini adalah masalah yang sangat penting di masa depan karena machine learning dapat digunakan untuk mengambil keputusan yang akurat. Ketika dihadapkan pada suatu masalah, peneliti statistik cenderung beralih ke rumusan matematika untuk memodelkan masalah tersebut. Teknik statistik dan machine learning masing-masing memiliki keunggulan, dan kedua bidang tersebut

saling melengkapi terutama dalam pengembangan teknologi big data.

DAFTAR PUSTAKA

Furqon, 2009. Statistika Terapan Untuk Penelitian. Bandung : Alfabeta

Harrell, Frank. 2015. Regression Modeling Strategies: With Applications To Linear Models, Logistic And Ordinal Regression, And Survival Analysis. Springer, 2015.

K. P. Murphy. 2012. Machine Learning: A Probabilistic Perspective. The MIT Press.

P. Dangeti, 2017. Statistics for Machine Learning. Mumbai:Packt.

Putra, Jan Wira Gotama. 2020. Pengenalan Konsep Pembelajaran Mesin dan Deep Learning Edisi 1.4, Tokyo Jepang.

Sugiyono. 2014. Metode Penelitian Pendidikan Pendekatan Kuantitatif, Kualitatif, dan R&D. Bandung: Alfabeta.