Technic Report

Visualize and Explore Data Using Breast Cancer Dataset

Tugas Untuk Memenuhi Mata Kuliah Machine Learning

Disusun Oleh:

Nama : Rai Barokah Utari

NIM : 1103200066

PROGRAM STUDI SI TEKNIK KOMPUTER

FAKULTAS TEKNIK ELEKTRO

TELKOM UNIVERSITY

BANDUNG

2023

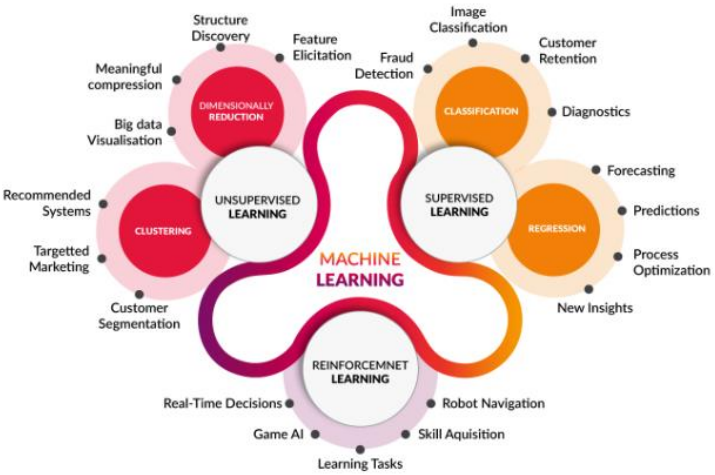
I. Mechine Learning

Teknologi machine learning (ML) adalah mesin yang dikembangkan untuk bisa belajar dengan sendirinya tanpa arahan dari penggunanya. Pembelajaran mesin dikembangkan berdasarkan disiplin ilmu lainnya seperti statistika, matematika dan data mining sehingga mesin dapat belajar dengan menganalisa data tanpa perlu di program ulang atau diperintah. Dalam hal ini machine learning memiliki kemampuan untuk memperoleh data yang ada dengan perintah ia sendiri. ML juga dapat mempelajari data yang ada dan data yang ia peroleh sehingga bisa melakukan tugas tertentu. Tugas yang dapat dilakukan oleh ML pun sangat beragam, tergantung dari apa yang ia pelajari.

Istilah machine learning pertama kali dikemukakan oleh beberapa ilmuwan matematika seperti Adrien Marie Legendre, Thomas Bayes dan Andrey Markov pada tahun 1920-an dengan mengemukakan dasar-dasar machine learning dan konsepnya. Sejak saat itu ML banyak yang mengembangkan. Salah satu contoh dari penerapan ML yang cukup terkenal adalah Deep Blue yang dibuat oleh IBM pada tahun 1996.

Sehingga, lebih singkatnya yaitu Machine learning adalah salah satu cabang ilmu kecerdasan buatan (artificial intelligence) yang mempelajari cara mesin (komputer) dapat belajar dari data dan membuat keputusan atau prediksi berdasarkan pengalaman atau pola yang terdapat dalam data tersebut, tanpa harus secara eksplisit diprogram secara khusus untuk setiap situasi. Dengan kata lain, machine learning memungkinkan komputer untuk belajar dari data yang diberikan dan meningkatkan kinerjanya seiring dengan bertambahnya jumlah data yang diproses. Machine learning digunakan di banyak bidang, seperti pengenalan suara dan gambar, pengolahan bahasa alami, dan analisis data. Ada beberapa jenis machine learning, termasuk supervised learning, unsupervised learning, dan reinforcement learning.

2. Model-model yang digunakan:



- Supervised Learning: Supervised Learning adalah sebuah pembelajaran dengan menggunakan Algoritma yang bertipe klasifikasi dengan kata lain datanya harus memiliki Label atau Tujuan akhir. Dalam prosesnya supervised learning memerlukan bantuan data yang dikumpulkan dari masa sebelumnya untuk melatih dan ditentukan sebuah model dari Algoritma yang dipilih.
- Unsupervised Learning: Pada unsupervised learning maka dalam proses tidak memerlukan data sebelumnya sebagai input. Dalam metode ini memungkinkan model untuk belajar sendiri menggunakan data yang telah diberikan.
- Reinforcement Learning: Reinforcement Learning adalah proses membuat model untuk belajar bagaimana membuat suatu keputusan. Teknik ini sebenarnya banyak diteliti pada machine learning karena Algoritma metode ini membantu model belajar berdasarkan umpan balik.

Selain yang diatas, ada beberapa model umum yang digunakan dalam machine learning, antara lain:

- Regresi: Model regresi digunakan untuk memprediksi nilai numerik berdasarkan variabel input yang diberikan. Contoh aplikasi regresi adalah prediksi harga rumah berdasarkan luas tanah, jumlah kamar tidur, dan lokasi.
- Klasifikasi: Model klasifikasi digunakan untuk memprediksi kategori atau label berdasarkan variabel input yang diberikan. Contoh aplikasi klasifikasi adalah pengenalan wajah, deteksi spam email, dan diagnosis penyakit.

- Clustering: Model clustering digunakan untuk mengelompokkan data berdasarkan kesamaan karakteristik. Contoh aplikasi clustering adalah segmentasi pasar dan klasifikasi dokumen.
- Association rule learning: Model association rule learning digunakan untuk menemukan hubungan antara item-item dalam dataset. Contoh aplikasi association rule learning adalah rekomendasi produk pada situs e-commerce dan analisis keranjang belanja.
- Deep learning: Model deep learning adalah model neural network yang memiliki banyak lapisan dan digunakan untuk mempelajari pola kompleks dalam data. Contoh aplikasi deep learning adalah pengenalan suara dan gambar, serta pengolahan bahasa alami.
- Reinforcement learning: Model reinforcement learning digunakan untuk mempelajari bagaimana agen (agent) dapat melakukan tindakan (action) untuk mencapai tujuan tertentu dalam lingkungan (environment) yang dinamis. Contoh aplikasi reinforcement learning adalah permainan video dan robotika.

I. Kanker Payudara

Kanker payudara adalah salah satu jenis kanker yang paling umum terjadi pada wanita di seluruh dunia. Kanker ini terjadi ketika sel-sel di dalam payudara tumbuh secara abnormal dan tidak terkontrol. Kanker payudara dapat terjadi pada pria, meskipun kasusnya jauh lebih jarang. Kanker payudara adalah suatu jenis tumor ganas yang berkembang pada sel-sel payudara. Kanker ini dapat tumbuh jika terjadi pertumbuhan yang abnormal dari sel-sel pada payudara. Sel-sel tersebut membelah diri lebih cepat dari sel normal, yang kemudian membentuk benjolan atau massa. Pada stadium yang lebih parah, sel-sel abnormal ini dapat menyebar melalui kelenjar getah bening ke organ tubuh lainnya. Kanker payudara dapat menyebabkan berbagai gejala seperti benjolan pada payudara, nyeri pada payudara, perubahan bentuk atau ukuran payudara, perubahan pada kulit payudara, dan perubahan pada puting susu. Meskipun demikian, tidak semua benjolan pada payudara adalah kanker payudara. Sebagian besar benjolan pada payudara adalah jinak atau non-kanker.

Pemeriksaan dini dan deteksi kanker payudara sejak dini dapat memperbaiki prognosis pasien dan memberikan peluang kesembuhan yang lebih baik. Pemeriksaan dapat dilakukan melalui pemeriksaan fisik, mamografi, USG, dan MRI. Selain itu, penelitian terus dilakukan untuk mengembangkan metode deteksi dan pengobatan

kanker payudara yang lebih efektif dan terjangkau. Pada technical report ini, akan dilakukan beberapa visualisasi data dan eksplorasi menggunakan beberapa teknik yang ada di dalam machine learning.

II. Pengumpulan Data

Tahap pertama yang dilakukan yaitu Import Library dan Load Dataset Pada bagian ini, dilakukan import beberapa library yang dibutuhkan untuk pengolahan data seperti numpy, pandas, matplotlib, seaborn, dan beberapa library dari sklearn, kemudian Eksplorasi Data Pada bagian ini, dilakukan beberapa eksplorasi data seperti melihat informasi umum dari data seperti jumlah baris dan kolom, tipe data dari masing-masing kolom, serta statistik deskriptif dari masing-masing kolom, Pemisahan Data pada bagian ini, dilakukan pemisahan data menjadi data train dan data test menggunakan train_test_split. Dan terakhir, kita dapat menggunakan heatmap untuk melihat tingkat korelasi antara setiap features yang ada di dalam dataset tersebut.

III. Sintaks yang digunakan

- Import numpy as np digunakan untuk mengimpor library NumPy dan memberikan alias atau nama lain np pada library tersebut. NumPy adalah library pada Python yang menyediakan struktur data array dan berbagai fungsi matematika untuk melakukan operasi pada array tersebut.
- Import pandas as pd digunakan untuk mengimpor library Pandas dan memberikan alias atau nama lain pd pada library tersebut. Pandas adalah library pada Python yang digunakan untuk melakukan manipulasi, analisis, dan visualisasi data dalam bentuk tabel atau data frame
- Import seaborn as sns digunakan untuk mengimpor library Seaborn dan memberikan alias atau nama lain sns pada library tersebut. Seaborn adalah library pada Python yang digunakan untuk membuat visualisasi data yang lebih menarik dan informatif dengan menyediakan berbagai jenis plot seperti scatter plot, line plot, bar plot, heatmap,dll.
- Import matplotlib.pyplot as plt digunakan untuk mengimpor library Matplotlib dan memberikan alias atau nama lain plt pada library tersebut. Matplotlib adalah library pada Python yang digunakan untuk membuat visualisasi data dalam berbagai bentuk plot

IV. Eksplorasi Data

Eksplorasi data dengan menggunakan teknik Random Forest untuk menggambarkan feature importances dari random forest classifier pada kanker payudara ini yang selanjutnya dapat kita lakukan visualisasi pada setiap features yang ada. Selanjutnya, kita juga dapat menentukan hasil dari akurasi skor, classification report, dan confusion matrix untuk model predictionnya yang selanjutnya dapat kita tampilkan dan mendapatkan hasilnya. Eksplorasi data terakhir, adalah dengan menggunakan self-training. Hal yang pertama kali kita lakukan adalah dengan melakukan split data untuk dipisah menjadi training sets dan testing sets. Selanjutnya, disini kita dapat mengubah data cancer tersebut menjadi Dataframe dan dapat kita lakukan plot untuk melihat distribusi dari target yang ada. teknik self-training classifier untuk mengklasifikasikan dataset breast cancer tersebut.

Teknik Self-Training yang dilakukan adalah dengan menggunakan semi-supervised learning di mana model akan diawali dengan beberapa data yang sudah dilabeli dan kemudian menggunakan data yang tidak dilabeli untuk melatih dirinya sendiri secara berulang sampai konvergen. Tujuan dari eksperimen ini adalah untuk membandingkan performa klasifikasi dan jumlah sampel yang dilabeli dengan mengatur parameter threshold pada Self-Training Classifier. Eksperimen akan ini menghasilkan dua grafik. Grafik pertama menunjukkan nilai akurasi rata-rata dari 3-fold cross validation dengan menggunakan threshold yang berbedabeda. Grafik ketiga menunjukkan jumlah iterasi yang diperlukan oleh model untuk konvergen dengan menggunakan threshold yang berbeda-beda.

Tahapan:

1. Mencari dataset dan import library

   Library yang digunakan adalah numpy untuk memanipulasi array atau matriks secara efisien, seaborn untuk memvisualisasikan data secara statistik, matplotlib untuk memvisualisasikan data dalam bentuk grafik atau plot, pandas untuk melakukan analisis data dan memanipulasi data secara efisien dengan cara membentuk data tersebut menjadi sebuah dataframe, dan beberapa dari sklearn untuk membangun model machine learning.

2. Visualisasi data

   Proses mempresentasikan data dalam bentuk grafik, diagram, atau plot agar mudah dipahami dan memberikan informasi yang lebih jelas dan komunikatif. Tujuan dari

visualisasi data adalah untuk mempermudah pembacaan, interpretasi, dan analisis data sehingga dapat membantu dalam pengambilan keputusan.

3. Eksplorasi data

   Proses mengumpulkan, menganalisis, dan memahami data yang dimiliki dengan tujuan untuk mendapatkan informasi yang berguna dan mendalam tentang data

V. Kesimpulan

Dapat disimpulkan bahwa dataset Breast Cancer Wisconsin (Diagnostic) dapat digunakan untuk memprediksi jenis tumor dengan akurasi yang cukup tinggi menggunakan dengan memilih fitur-fitur yang tepat dan melakukan transformasi data yang diperlukan. Visualisasi data dilakukan menggunakan sns.swarmplot dan heatmap, sedangkan eksplorasi data dilakukan dengan membagi data menjadi data test dan data latih pada algoritma tree decision dan random forest, serta self-training. Selain itu, pada eksplorasi data juga ditemukan beberapa hal yang perlu diperhatikan dalam pemrosesan data, seperti normalisasi data dan deteksi outliers.

Referensi:

https://www.dicoding.com/blog/machine-learning-adalah/
https://dqlab.id/5-metode-machine-learning-yang-sering-digunakan-data-engineer
https://chat.openai.com/