TECHNICAL REPORT

VISUALIZE AND EXPLORE DATA USING BREAST CANCER DATASET

TUGAS UNTUK MEMENUHI MATA KULIAH

MACHINE LEARNING

Oleh:

Muhammad Tharreq An Nahl

1103204040

PROGRAM STUDI SI TEKNIK KOMPUTER

FAKULTAS TEKNIK ELEKTRO

TELKOM UNIVERSITY

BANDUNG

2023

A. Pengertian Machine Learning

Machine learning merupakan suatu cabang ilmu dari kecerdasan buatan (artificial intelligence) yang berkaitan dengan pengembangan dari algoritma dan teknik untuk memungkinkan suatu mesin atau komputer dapat belajar dari data yang diberikan dan melakukan pengambilan keputusan secara otomatis. Dalam konteks machine learning, "belajar" dapat mengacu pada kemampuan mesin untuk mengenali pola dan hubungan dalam data yang diberikan, dan menggunakan informasi ini untuk memperbaiki kinerja atau hasilnya.

Dalam prakteknya, machine learning biasanya melibatkan pemrosesan data secara besar-besaran, baik itu data terstruktur (seperti data dari basis data atau tabel Excel) maupun data tidak terstruktur (seperti audio, gambar, atau teks). Proses machine learning terdiri dari beberapa tahap, mulai dari persiapan data, pemilihan algoritma machine learning yang sesuai, pelatihan model, evaluasi kinerja model, hingga penggunaan model untuk melakukan prediksi atau pengambilan keputusan.

B. Machine Learning Model

Terdapat beberapa model umum yang ada pada machine learning jika kita merujuk pada pendekatan atau teknik yang digunakan untuk membangun model yang dapat mempelajari pola dan mengambil keputusan atau prediksi berdasarkan data yang diberikan. Di bawah ini merupakan beberapa model umum yang digunakan pada machine learning:

1. Regresi: Model regresi digunakan untuk memprediksi nilai numerik berdasarkan satu atau beberapa variabel bebas. Misalnya, regresi linier digunakan untuk memprediksi nilai variabel dependen berdasarkan satu variabel independen.
2. Klasifikasi: Model klasifikasi digunakan untuk memprediksi label atau kelas tertentu berdasarkan fitur atau atribut yang diberikan. Contoh aplikasi klasifikasi adalah deteksi spam di email, klasifikasi citra, atau identifikasi jenis tumor berdasarkan gambar medis.
3. Pengelompokan (*clustering*): Model pengelompokan digunakan untuk mengelompokkan data berdasarkan kesamaan fitur atau atribut. Model pengelompokan umumnya digunakan untuk eksplorasi data atau untuk mempermudah analisis data yang besar.

4. Jaringan Saraf Tiruan (Artificial Neural Networks): Jaringan Saraf Tiruan (ANN) adalah model machine learning yang terinspirasi dari cara kerja otak manusia. ANN terdiri dari banyak "neuron" yang saling terhubung dan digunakan untuk memproses data masukan dan menghasilkan keluaran.
5. Pohon Keputusan (Decision Tree): Model pohon keputusan digunakan untuk membuat keputusan berdasarkan serangkaian aturan yang disusun dalam bentuk pohon. Pohon keputusan biasanya digunakan untuk memprediksi kelas atau label berdasarkan fitur atau atribut yang diberikan.
6. Metode Ensemble: Metode ensemble adalah teknik yang menggabungkan beberapa model machine learning menjadi satu model yang lebih kuat. Contoh dari metode ensemble adalah Random Forest dan Gradient Boosting.

C. Machine Learning Model untuk Pengklasifikasian Tugas

Ada banyak model machine learning yang dapat digunakan untuk tugas klasifikasi, tetapi berikut ini adalah tiga model teratas yang sering digunakan dan terbukti memberikan hasil yang baik:

1. Support Vector Machines (SVM): SVM adalah model pembelajaran mesin yang digunakan untuk klasifikasi biner atau multikelas. SVM berusaha untuk menemukan "*hyperplane*" yang memaksimalkan jarak antara kelas yang berbeda pada data. SVM sangat berguna dalam mengklasifikasikan data yang kompleks, terutama ketika terdapat banyak fitur pada data.
2. Random Forest: Random Forest adalah model pembelajaran mesin ensemble yang terdiri dari banyak pohon keputusan yang digabungkan menjadi satu model yang lebih kuat. Random Forest sangat berguna dalam mengklasifikasikan data yang beragam dan kompleks, terutama ketika terdapat banyak fitur pada data. Random Forest juga dapat mengatasi masalah overfitting pada model.
3. Naive Bayes: Naive Bayes adalah model pembelajaran mesin yang sederhana dan efektif untuk klasifikasi. Naive Bayes menghitung probabilitas kelas berdasarkan probabilitas fitur dalam data. Naive Bayes sangat cepat dalam melakukan klasifikasi, bahkan pada data yang sangat besar. Naive Bayes juga terbukti memberikan hasil yang baik pada data teks dan data kategori.

D. Dataset Publik yang tersedia untuk Breast Cancer

Terdapat beberapa dataset publik yang tersedia untuk kanker payudara. Beberapa di antaranya adalah:

1. Breast Cancer Wisconsin (Diagnostic) Dataset: Dataset ini tersedia di UCI Machine Learning Repository dan terdiri dari 569 sampel sel kanker payudara, dengan 30 fitur numerik yang diambil dari citra digital biopsi payudara. Dataset ini digunakan untuk memprediksi apakah sel kanker bersifat ganas atau jinak.
2. Breast Cancer Wisconsin (Original) Dataset: Dataset ini juga tersedia di UCI Machine Learning Repository dan terdiri dari 699 sampel sel kanker payudara, dengan 10 fitur numerik yang diambil dari citra digital biopsi payudara. Dataset ini digunakan untuk memprediksi apakah sel kanker bersifat ganas atau jinak.
3. Wisconsin Diagnostic Breast Cancer (WDBC) Dataset: Dataset ini tersedia di Kaggle dan terdiri dari 569 sampel sel kanker payudara, dengan 30 fitur numerik yang diambil dari citra digital biopsi payudara. Dataset ini digunakan untuk memprediksi apakah sel kanker bersifat ganas atau jinak.
4. The Cancer Imaging Archive (TCIA): TCIA adalah sebuah repository data medis yang berisi berbagai jenis data medis, termasuk data citra payudara. TCIA menyediakan beberapa dataset citra payudara, seperti dataset Mammography Image Analysis Society (MIAS) dan dataset Digital Database for Screening Mammography (DDSM), yang dapat digunakan untuk melakukan penelitian kanker payudara.

Semua dataset ini dapat digunakan untuk melakukan penelitian dan pengembangan model machine learning untuk kanker payudara. Namun, sebelum menggunakannya, penting untuk memahami karakteristik data dan sumber dataset untuk memastikan bahwa data yang digunakan sesuai dengan kebutuhan penelitian.

E. Penjelasan Konten dari Breast Cancer Dataset

Kumpulan data breast cancer dataset merujuk pada dua dataset yang disediakan di UCI Machine Learning Repository, yaitu Breast Cancer Wisconsin (Diagnostic) Dataset dan Breast Cancer Wisconsin (Original) Dataset. Kedua dataset tersebut memuat informasi mengenai sel kanker payudara, dengan tujuan untuk memprediksi apakah sel kanker bersifat ganas atau jinak.

Breast Cancer Wisconsin (Diagnostic) Dataset terdiri dari 569 sampel sel kanker payudara, di mana setiap sampel direpresentasikan oleh 30 fitur numerik yang diambil dari citra digital biopsi payudara. Fitur-fitur ini mencakup informasi mengenai ukuran sel, bentuk, tekstur, dan parameter lain yang dihitung dari citra sel kanker. Setiap sampel dilabeli sebagai ganas (malignant) atau jinak (benign), sehingga dapat digunakan untuk membangun model klasifikasi untuk memprediksi apakah sel kanker bersifat ganas atau jinak.

Breast Cancer Wisconsin (Original) Dataset juga terdiri dari sampel sel kanker payudara, tetapi dengan jumlah sampel yang lebih besar, yaitu 699 sampel. Setiap sampel direpresentasikan oleh 10 fitur numerik yang diambil dari citra digital biopsi payudara. Fitur-fitur ini mencakup informasi mengenai ukuran sel, bentuk, dan parameter lain yang dihitung dari citra sel kanker. Setiap sampel juga dilabeli sebagai ganas atau jinak.

Kumpulan data breast cancer dataset sangat berguna dalam pengembangan model machine learning untuk deteksi kanker payudara, khususnya pada tugas klasifikasi ganas/jinak. Dataset ini telah digunakan dalam banyak penelitian di bidang kesehatan dan ilmu komputer, dan menjadi salah satu dataset standar untuk penelitian di bidang ini.

F. Pendahuluan

Kanker Payudara merupakan salah satu jenis kanker yang paling umum pada kalangan wanita di seluruh dunia. Pendeteksian dan diagnosis dini dari kanker payudara merupakan suatu faktor penting yang dapat meningkatkan tingkat kelangsungan hidup dari kalangan wanita untuk dapat mencegah mereka dari kematian. Oleh karena itu, dengan adanya kegiatan eksplorasi dan visualisasi mengenai kanker payudara ini dapat membantu para peneliti dan dokter di seluruh dunia untuk dapat mengetahui faktor-faktor apa saja yang dapat mempengaruhi terjadinya perkembangan kanker payudara ini.

Pada technical report ini, akan dilakukan beberapa visualisasi data dan eksplorasi menggunakan beberapa teknik yang ada di dalam machine learning. Untuk dataset yang diambil akan memiliki beberapa informasi mengenai parameter-parameter yang akan digunakan untuk memprediksi beberapa hal dari kanker payudara tersebut.

G. Visualisasi Data

Hal yang pertama kali akan dilakukan dalam kegiatan visualisasi dan eksplorasi data kanker payudara ini adalah melakukan *import* beberapa libraries dan module yang akan digunakan di dalam python seperti NumPy, Pandas, Seaborn, Scikit-learn, dll. Setelah itu, kita dapat melakukan load data breast cancer yang terdapat di dalam dataset scikit-learn dan akan disimpan di dalam variabel cancer.

Selanjutnya, kita dapat melakukan visualisasi data pada feature pertama yang ada didalam dataset (‘mean radius’) dengan menggunakan ‘sns.histplot()’. Jika sudah, kita juga dapat melakukan visualisasi data dari persebaran target variabel breast cancer tersebut. Dan terakhir, kita dapat menggunakan heatmap untuk melihat tingkat korelasi antara setiap features yang ada di dalam dataset tersebut.

H. Explorasi Data

Eksplorasi data pertama yang dilakukan adalah dengan menggunakan teknik Decision Tree untuk menunjukkan penerapan dari pemangkasan kompleksitas biaya dalam melakukan pengklasifikasian dengan menggunakan pohon keputusan. Hal yang pertama kali dilakukan adalah melakukan split breast cancer dataset menjadi training sets dan testing sets dengan menggunakan ‘train_test_split’. Selanjutnya, dilakukan ‘cost_complexity_pruning_path’ untuk mengkalkulasi optimal value dari ‘ccp_alpha’ parameter yang akan mengontrol tingkat kompleksitas dari decision tree. Pada akhirnya, output yang dikeluarkan pada bagian ini akan mencetak jumlah node pada pohon terakhir yang terdapat didalam list ‘clfs’ dan nilai optimal dari ‘clfs’ dan nilai dari ‘ccp_alpha’.

Selanjutnya, kita dapat melakukan visualisasi data tersebut dengan menggunakan libraries seaborn untuk memperlihatkan gambar dari total jumlah nodes vs alpha dan nilai depth vs alpha. Selain itu, kita juga dapat melihat hasil visualisasi nilai Accuracy vs Apha yang digunakan untuk training and testing sets. Dan terakhir kita dapat memvisualisasikan hasil decision tree tersebut dengan menggunakan fungsi tree_plot_tree.

Eksplorasi data yang kedua adalah 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.

Selanjutnya, kita dapat melakukan 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 berbeda-beda. Grafik ketiga menunjukkan jumlah iterasi yang diperlukan oleh model untuk konvergen dengan menggunakan threshold yang berbeda-beda.