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. 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.

B. 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.

C. 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.