Laporan Teknis

Visualisasi dan Eksplorasi Data Kanker Payudara Menggunakan *Tree Decision, Random Forest,* dan *Self-Training*

Laporan ini membahas hasil dari visualisasi dan eksplorasi data kanker payudara dengan menggunakan algoritma *tree decision*, *random forest*, dan *self-training*. Tujuan dari penelitian ini adalah untuk menganalisis dan memahami hubungan antara faktor risiko tertentu dan kemungkinan terjadinya kanker payudara. Data yang digunakan dalam penelitian ini diperoleh dari situs *kaggle.com* yang terkait dengan kanker payudara. Hasil dari analisis data menunjukkan bahwa algoritma *tree decision*, *random forest*, dan *self-training* dapat digunakan untuk memprediksi risiko kanker payudara dengan akurasi yang tinggi. Semoga laporan ini dapat memberikan pemahaman yang lebih baik terhadap faktor risiko kanker payudara dan kontribusi yang positif dalam upaya pencegahan dan pengobatan kanker payudara.

Tahap 1: Mencari dataset dan import library yang digunakan

Dataset yang digunakan sudah ditentukan, yaitu menggunakan dataset *breast cancer* menggunakan *API* dari *scikit-learn*. Hanya saja terdapat data kolom yang hilang, sehingga pada penugasan ini menggunakan dataset *breast cancer* dari *kaggle.com* yang isinya mirip-mirip dan lebih lengkap, sehingga memudahkan dalam melakukan visualisasi data. Tautannya adalah *https://www.kaggle.com/datasets/nancyalaswad90/breast-cancer-dataset*.

Library yang digunakan adalah *numpy, seaborn, matplotlib, pandas, dan beberapa dari sklearn*. Yang paling banyak digunakan pada proyek ini adalah *pandas* karena dataset yang berupa *csv* ini disimpan secara lokal, kemudian diubah ke bentuk *dataframe* supaya mudah pengolahannya. Kemudian *matplotlib* dan *seaborn* untuk menampilkan grafik.

Tahap 2: Visualisasi data

Dataset yang sudah diubah ke bentuk *dataframe*, kemudian lakukan visualisasi. Dari yang paling mudah, mengelompokkan tumor yang ganas dengan yang jinak. Kemudian melakukan visualisasi data yang lebih mendetil lagi menggunakan *sns.swarmplot*. Dari grafik yang dihasilkan dapat terlihat jelas perbatasan antara tumor ganas dengan tumor jinaknya berdasarkan data lainnya. Selain itu juga, pada proyek ini terdapat grafik *heatmap* yang menampilkan korelasi data satu dengan yang lainnya.

Tahap 3: Eksplorasi data

Dari ketiga metode, untuk yang *Tree Decision* dan *Random Forest* sudah dibagi data tes dan data latihnya dengan perbandingan 1:4. Setelah itu program akan menampilkan akurasinya. Sedangkan untuk *self-training* menggunakan *gamma* sebesar 0,001, kemudian program akan menampilkan akurasi terhadap jumlah iterasinya dalam bentuk grafik.

Kesimpulan

Visualisasi dan eksplorasi data kanker payudara menggunakan algoritma *tree decision, random forest,* dan *self-training* dapat digunakan untuk menganalisis dan memahami hubungan antara faktor risiko tertentu dan kemungkinan terjadinya kanker payudara. Dalam penelitian ini, data yang digunakan berasal dari situs kaggle.com dan telah diolah menggunakan *library numpy, seaborn, matplotlib,* dan *pandas*. 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*. Hasil analisis data menunjukkan bahwa algoritma *tree decision, random forest,* dan *self-training* dapat digunakan untuk memprediksi risiko kanker payudara dengan akurasi yang tinggi.