PRAKTIKUM 3
FEATURE ENGINEERING
IF4074 PEMBELAJARAN MESIN LANJUT

Disusun Oleh :

13520041 Ilham Pratama

13520042 Jeremy S.O.N. Simbolon

PROGRAM STUDI INFORMATIKA
SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA
INSTITUT TEKNOLOGI BANDUNG
2023

A. Desain Experiment

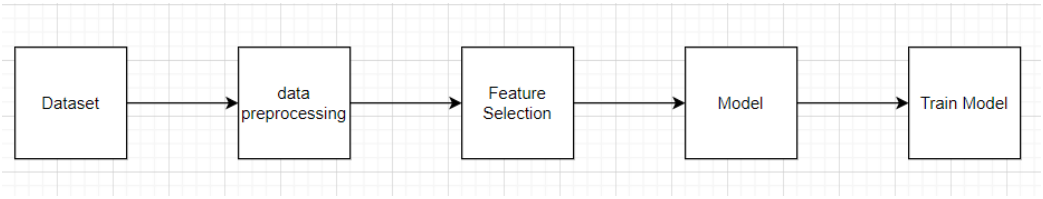
a. Dataset

Dataset yang digunakan pada praktikum ini adalah dataset Titanic. Dataset ini digunakan untuk melakukan klasifikasi berdasarkan atribut pada dataset untuk menentukan apakah seseorang selamat atau tidak. Atribut-Atribut yang ada pada dataset ini adalah:

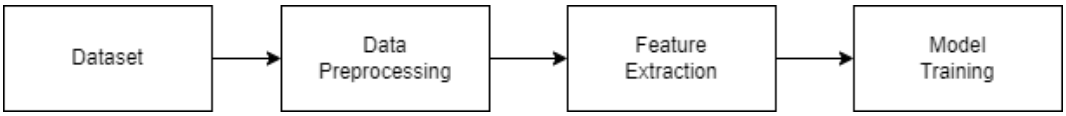
1) Survival, yaitu target atau label dari proses klasifikasi, ada 2 nilai yaitu 0 (tidak selamat) dan 1 (Selamat)
2) Pclass, yaitu kelas tiket, ada 3 nilai yaitu 1 (1st), 2 (2nd), dan 3 (3rd)
3) Sex, yaitu jenis kelamin
4) Age, yaitu umur penumpang dalam tahun
5) Sibsp, yaitu jumlah saudara yang ada pada kapal
6) Parch, yaitu jumlah orang tua atau anak yang ada pada kapal
7) Ticket, yaitu nomor tiker
8) Fare, yaitu biaya tiket penumpang
9) Cabin, yaitu nomor kabin
10) Embarked, yaitu pelabuhan keberangkatan, ada 3 nilai yaitu C (Cherbourg), Q (Queenstown), dan S (Southampton)

b. Schema Diagram

i. Feature Selection



ii. Feature Extraction



c. Feature Selection

Feature selection yang digunakan pada atribut ini adalah f_classif dan chi square. Alasan pemilihan fitur ini adalah karena kedua fitur ini digunakan untuk menentukan korelasi atribut yang dirancang dengan variabel kategoris lainnya pada dataset, Nilai yang tinggi menunjukkan ada korelasi yang kuat antara 2 atribut.

d. Feature Extraction

Metode *feature extraction* yang dipilih pada eksperimen ini adalah metode *linear discriminant analysis* (LDA). Metode ini dipilih untuk meminimumkan variansi antardata dengan kelas yang sama. Akan dibandingkan akurasi model yang dilatih menggunakan dataset yang dikenakan dan tidak dikenakan LDA.

B. Langkah-Langkah Experiment

a. Feature Selection

Langkah-langkah yang dilakukan pada Feature Selection adalah

1) Melakukan *preprocessing* terhadap data (dihasilkan 7 kolom atribut)
    a) Melakukan penanganan terhadap *null value*
    b) Melakukan *categorical encoding* terhadap kolom sex dan embarked
    c) Melakukan binning terhadap kolom Fare yang dibagi menjadi 2 nilai
    d) Menghilangkan beberapa kolom yang tidak diperlukan dalam proses klasifikasi
2) Feature Selection
    a) Melakukan seleksi kolom dengan metode f_classif dengan hanya mengambil 5 kolom
    b) Melakukan seleksi kolom dengan metode chi2 dengan hanya mengambil 5 kolom
3) Membuat Model
    a) Membuat model berupa model klasifikasi dengan menggunakan logistic regression
4) Melakukan *train* terhadap model, untuk train dataset dibagi menjadi 80% untuk proses train dan 20% untuk test
    a) Melakukan train terhadap model dengan menggunakan data yang belum diseleksi
    b) Melakukan train terhadap model dengan menggunakan data yang sudah diseleksi dengan metode f_classif
    c) Melakukan train terhadap model dengan menggunakan data yang sudah diseleksi dengan metode chi2

b. Feature Extraction

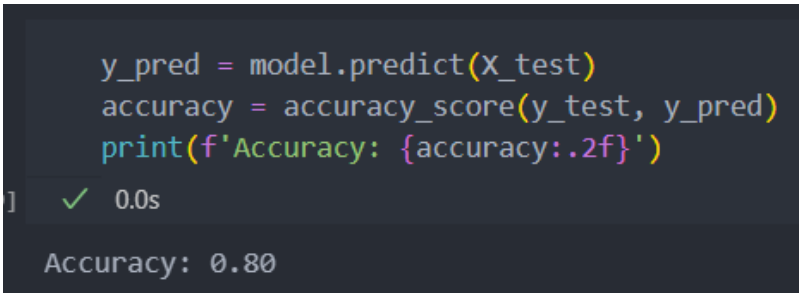
Langkah-langkah eksperimen *feature extraction* adalah berikut.

1) Melakukan *preprocessing* terhadap data.
    a) Melakukan penanganan terhadap data yang hilang.
    b) Membuang kolom yang tidak diperlukan.
    c) Melakukan *label encoding* pada kolom “Sex” dan “Embarked”.
    d) Melakukan normalisasi kolom menggunakan MinMaxScaler.
2) Melakukan *feature extraction* menggunakan LinearDiscriminantAnalysis.
3) Melatih model klasifikasi.
    a) Membagi dataset menjadi *train* dan *test set* dengan perbandingan 0.8:0.2.
    b) Melatih dan menghitung akurasi model LogisticRegression menggunakan train set yang tidak dikenakan *linear discriminant analysis*.
    c) Melatih dan menghitung akurasi model LogisticRegression menggunakan train set yang dikenakan *linear discriminant analysis*.

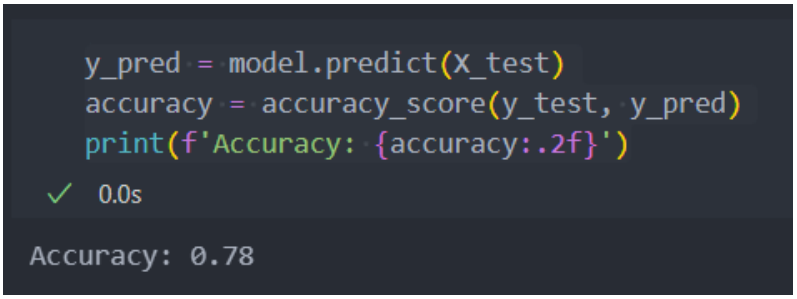
C. Hasil Experiment

a. Feature Selection

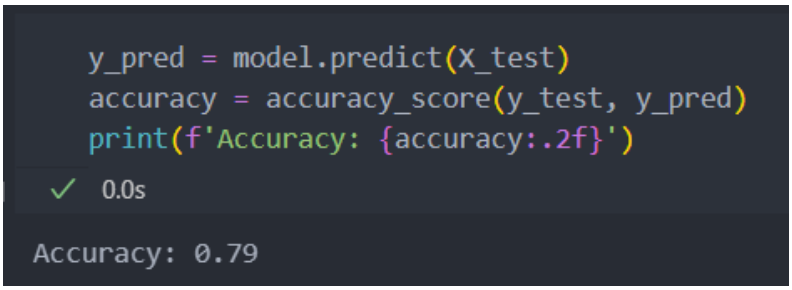
i. Dataset yang belum diseleksi



ii. Dataset yang sudah diseleksi dengan metode f_classif



iii. Dataset yang sudah diseleksi dengan metode chi2



b. Feature Extraction

i. Dataset yang belum dikenakan LDA

```
# To make the non-LDA model, we shall use the following pipeline:
# MinMaxScaler  -> LogisticRegression
# The MinMaxScaler is needed to ensure the higher value features wouldn't
# influence the model much more that the other features.
# The LogisticRegression is used as the classifier for the model.
pipe = make_pipeline(MinMaxScaler(), LogisticRegression())
pipe.fit(X_train, y_train)
pipe.score(X_test, y_test)
```

Executed at 2023.11.24 21:28:56 in 92ms

```
0.8324022346368715
```

ii. Dataset yang telah dikenakan LDA

```
# To make the LDA model, we shall use the following pipeline:
# MinMaxScaler -> LinearDiscriminantAnalysis -> LogisticRegression
# The MinMaxScaler is needed to ensure the higher value features wouldn't
# influence the model much more that the other features.
# The LinearDiscriminantAnalysis is used to project the dataset onto a
# lower-dimensional space, extracting features.
# The LogisticRegression is used as the classifier for the model.
lda_pipe = make_pipeline(MinMaxScaler(), LinearDiscriminantAnalysis(), LogisticRegression())
lda_pipe.fit(X_train, y_train)
lda_pipe.score(X_test, y_test)
```

Executed at 2023.11.24 21:28:56 in 95ms

```
0.8379888268156425
```

D. Analisis

a. Feature Selection

Berdasarkan pengujian yang dilakukan, didapatkan hasil yang tidak terlalu berbeda jauh dari segi accuracy dari model yang sudah dilatih. Dataset yang tanpa dilakukan proses seleksi memiliki accuracy yang lebih tinggi yaitu 80%. Berbeda dengan hasil dengan dataset yang sudah diseleksi kolomnya. Hasilnya sedikit berbeda dari hasil sebelumnya karena kurangnya jumlah kolom. Namun, perbedaannya tidak terlalu signifikan, yaitu hanya sekitar 1-2%. Hal ini menunjukkan bahwa proses seleksi kolom tidak terlalu mempengaruhi akurasi model.

b. Feature Extraction

Berdasarkan eksperimen yang dilakukan, dapat diamati adanya peningkatan akurasi model setelah dataset dikenakan *linear discriminant analysis*. Model yang dilatih menggunakan dataset yang tidak dikenakan LDA memiliki akurasi 0.8324, sedangkan model yang dilatih menggunakan dataset yang dikenakan LDA memiliki akurasi 0.8380. Dapat disimpulkan proses *feature extraction* memiliki pengaruh kecil terhadap akurasi model.

E. Kesimpulan

a. Feature Selection

Berdasarkan hasil pengujian dan analisis yang sudah dibahas sebelumnya, dapat ditarik suatu kesimpulan, yaitu pada skema yang digunakan feature selection tidak terlalu mempengaruhi akurasi dari model.

b. Feature Extraction

Berdasarkan hasil eksperimen, dapat disimpulkan proses *feature extraction* dapat meningkatkan akurasi model, walaupun pengaruh yang dapat diobservasi pada eksperimen ini bersifat minimal.

F. Lesson Learned / Improvement

a. Feature Selection

- Tidak semua dataset perlu dilakukan feature selection. Bisa dilihat pada bagian feature selection, setelah dilakukan proses feature selection accuracy dari model berkurang sekitar 1% - 2%.
- Untuk Improvement bisa dilakukan dengan menggunakan metode lain untuk feature selectionnya, seperti menggunakan metode wrapper dan embedded

b. Feature Extraction

- Perlu dipelajari lebih lanjut alternatif metode *feature extraction* yang dipilih, seperti *principal component analysis* dan penggunaan *autoencoder*, beserta pengaruhnya terhadap unjuk kerja model.

G. Pembagian Tugas

NIM	Nama	Tugas
13520041	Ilham Pratama	● Feature Selection ● Laporan
13520042	Jeremy S.O.N Simbolon	● Feature Extraction ● Laporan