Praktikum 3: *Feature Engineering*

IF4074 Pembelajaran Mesin Lanjut

Dipersiapkan oleh:

Nama	NIM
M Dwinta Harits Cahyana	13519041
Aisyah Farras Aqila	13519054
Alvin Rizqi Alfisyahrin	13519126

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

2022

Daftar Isi

1 Desain Eksperimen

1.1 Dataset

Pada praktikum ini, kami menggunakan dataset Titanic yang terbagi menjadi 2 jenis dataset, training dan testing. Dataset Titanic ini merupakan kumpulan data penumpang kapal Titanic dan menentukan apakah penumpang tersebut *survive* atau tidak. Dataset ini memiliki 12 fitur, yaitu:

1. PassengerId: Merupakan ID unik yang dimiliki oleh setiap penumpang
2. Survived: Variabel target yang akan di-predict
3. Pclass: Merupakan status sosial ekonomi penumpang dan merupakan fitur ordinal kategori yang memiliki 3 nilai unik, yaitu 1 untuk *upper class,* 2 untuk *middle class*, 3 untuk *lower class*.
4. Name: Nama dari penumpang
5. Sex: Jenis kelamin penumpang
6. Age: Umur penumpang
7. SibSp: jumlah saudara dan pasangan penumpang
8. Parch: jumlah orangtua dan anak penumpang
9. Ticket: nomor tiket penumpang
10. Fare: Tarif penumpang
11. Cabin: Nomor Cabin penumpang
12. Embarked: Merupakan pelabuhan embarkasi penumpang dan merupakan fitur kategorikal kategori yang memiliki 3 nilai unik, yaitu C untuk *Cherbourg,* Q untuk *QueensTown*, S untuk *Southampton*.

1.2 Tujuan Eksperimen

Tujuan dari eksperimen adalah menentukan algoritma Feature Engineering yang menghasilkan prediksi paling baik.

1.3 Variabel pada Eksperimen

Variabel-variabel pada eksperimen ini adalah:

1. Variabel bebas
   Variabel bebas pada eksperimen ini adalah algoritma Feature Engineering yang digunakan. Akan dibandingkan dua algoritma Feature Engineering, yaitu *feature selection* dengan metode filter dan *feature extraction* dengan metode LDA.
2. Variabel tetap
   Variabel tetap pada eksperimen ini adalah dataset yang digunakan, termasuk *task-task* yang dilakukan pada Feature Engineering sebelum melakukan *feature selection*/*feature extraction* (*data cleansing*, *data normalization*, *data formatting*).
3. Variabel kontrol
   Variabel kontrol pada eksperimen ini adalah algoritma model yang digunakan untuk melakukan prediksi, yaitu RandomForest.

1.4 Metode Evaluasi

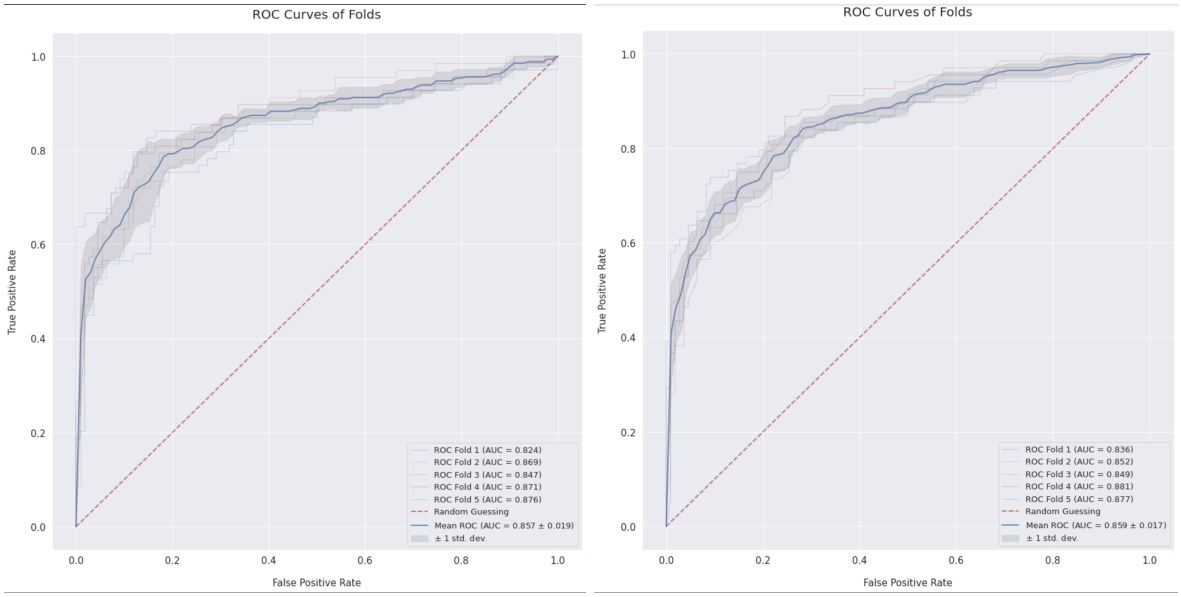
Kami melakukan evaluasi menggunakan metode *ROC - AUC Curve*.

2 Langkah Eksperimen

Berikut merupakan skema/langkah-langkah yang dilakukan pada eksperimen:

1. Melakukan *split* dataset untuk *train* dan *test*
2. Melakukan *data cleansing*, *data normalization*, dan *data formatting*
3. Membagi dataset menjadi dua, menggunakan algoritma Feature Engineering yang berbeda untuk kedua dataset tersebut
4. Melakukan training untuk kedua dataset
5. Membandingkan hasil training untuk kedua dataset menggunakan ROC - AUC Curve

3 Hasil Eksperimen



(a) (b)

Gambar 1. Hasil Eksperimen, (a) ROC Curve untuk Feature Selection dengan metode Filter, (b) ROC Curve untuk Feature Extraction dengan metode LDA

Dapat dilihat bahwa metode (a) menghasilkan AUC Score bernilai 0.857 dengan error 0.019 untuk Mean ROC, sedangkan metode (b) menghasilkan AUC Score bernilai 0.859 dengan error 0.017 untuk Mean ROC.

4 Analisis Hasil Eksperimen

Kami memilih RandomForest sebagai model training karena beberapa faktor, yaitu:

1. *Versatile*
   Dapat digunakan untuk dataset yang memiliki binary features, categorical features, dan numerical features.

2. *Paralellizable*
   Kita dapat membagi proses untuk dijalankan beberapa mesin sehingga menghasilkan waktu komputasi yang lebih cepat.
3. *Best with High Dimensionality*
   Random Forest sangat bagus dengan data berdimensi tinggi karena metode tersebut bekerja dengan subset dari data.

Namun terdapat kekurangan dengan metode ini, yaitu kecenderungan untuk overfit, sehingga kita perlu melakukan *hyperparameter tuning*. *Hyperparameter tuning* yang kami lakukan adalah criterion='gini', n_estimators=1100, max_depth=5, min_samples_split=4, min_samples_leaf=5, max_features='auto', oob_score=True, random_state=SEED, n_jobs=-1, verbose=1. Selain itu, kami juga menggunakan StratifiedKFold untuk stratifikasi variabel target. *Folds* dibuat dengan mempertahankan persentase sampel untuk setiap kelas dalam variabel target (Survived).

Lalu, kami melakukan evaluasi menggunakan ROC - AUC Curve. Kurva AUC - ROC adalah pengukuran kinerja untuk masalah klasifikasi pada berbagai pengaturan *threshold*. ROC adalah kurva probabilitas dan AUC merepresentasikan tingkat *separability*. Kurva ini memberitahu seberapa baik model mampu membedakan antar kelas. Semakin tinggi AUC, semakin baik model dalam memprediksi kelas 0 sebagai 0 dan kelas 1 sebagai 1.

Dapat dilihat bahwa feature extraction menggunakan metode LDA memiliki nilai AUC yang lebih baik dibanding feature selection menggunakan metode Filter, sehingga metode LDA dapat membedakan antar kelas lebih baik.

5 Kesimpulan

Kelompok kami menggunakan dataset Titanic dan telah dilakukan eksperimen terkait penggunaan algoritma *feature selection* dengan metode filter dan *feature extraction* dengan metode LDA. Kedua algoritma ini dibandingkan performanya menggunakan model RandomForest dengan metode evaluasi *ROC - AUC Curve*. Hasil eksperimen tersebut menunjukkan bahwa algoritma *feature extraction* dengan metode LDA menghasilkan performa yang lebih baik karena nilai AUC nya yang lebih tinggi dibanding algoritma *feature selection* dengan metode filter.

6 Lesson Learned

1. Kolom ‘Cabin’ dapat dieksplorasi lebih lanjut daripada di-*drop* langsung.
2. Dapat dilakukan eksplorasi lebih lanjut terhadap korelasi antar fitur yang ada di dataset untuk membuat fitur-fitur baru, sebagai contoh membuat kolom frekuensi ticket yang merupakan penghitungan kemunculan nomor tiket di dataset
3. Dapat dilakukan *target encoding* dengan memperhatikan nama dari setiap penumpang untuk membuat fitur baru seperti apakah penumpang tersebut sudah menikah atau tidak berdasarkan title yang mereka punya.

7 Pembagian Tugas

Nama	NIM	Tugas
Mohammad Dwinta Harits Cahyana	13519041	-
Aisyah Farras Aqila	13519054	• Feature Selection • Feature Extraction • Modelling • Laporan
Alvin Rizqi Alfisyahrin	13519126	• Exploratory Data Analysis • Feature Engineering Preprocess • Modelling • Laporan