Prediksi Hujan di Denpasar menggunakan *Denpasar Weather Data*

Laporan Praktikum IF3270 Pembelajaran Mesin

Kelas 3

Gerald Abraham Sianturi	13520138
Daffa Romyz Aufa	13520162

SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA

TEKNIK INFORMATIKA

2023

1. Hasil analisis data

1.1. *Duplicate value*

```
# Duplicate value
countOfDuplicateValue = data.duplicated().value_counts()[1]
print(f"Banyak nilai duplikat: {countOfDuplicateValue}")
```

```
Banyak nilai duplikat: 7253
```

Berdasarkan hasil yang diperoleh, terdapat 7.253 *rows* yang merupakan *row* dengan nilai duplikat (nilai yang identik dengan *row* lain).

1.2. *Missing value*

```
# Missing value
numOfRows = data.shape[0]

listOfCountMissingVal = data.isna().sum().values
for i, col in enumerate(data):
    numOfMissingValue = listOfCountMissingVal[i]
    proportionOfMissingValue = round(numOfMissingValue / numOfRows * 100, 2)
    print(f"{col}: {numOfMissingValue} ({proportionOfMissingValue} %)")
```

```
hour: 0 (0.0 %)
temp: 0 (0.0 %)
temp_min: 0 (0.0 %)
temp_max: 0 (0.0 %)
pressure: 0 (0.0 %)
humidity: 0 (0.0 %)
wind_speed: 0 (0.0 %)
wind_deg: 0 (0.0 %)
raining: 0 (0.0 %)
```

Berdasarkan hasil yang diperoleh, tidak ditemukan *cell* yang merupakan *missing value*.

1.3. *Outlier*

```
# Outlier
numericFeatNoOutlier = []
for feature in numericFeatures:
  Q1 = X[feature].quantile(0.25)
  Q3 = X[feature].quantile(0.75)
  IQR = round(Q3 - Q1, 2)
  countOutlier = ((X[feature] < (Q1 - 1.5 * IQR)) | (X[feature] > (Q3
+ 1.5 * IQR))).sum()
  proportionOutlier = round(countOutlier / numOfRows * 100, 2)
  print(f"{feature}: {countOutlier} ({proportionOutlier} %)")
  if(countOutlier == 0):
    numericFeatNoOutlier.append(feature)

numericFeatWithOutlier = list(set(numericFeatures) - set(numericFeatNoOutlier))
```

```
hour: 0 (0.0 %)
temp: 1458 (0.55 %)
temp_min: 1716 (0.65 %)
temp_max: 547 (0.21 %)
pressure: 1067 (0.4 %)
humidity: 231 (0.09 %)
wind_speed: 3439 (1.3 %)
wind_deg: 0 (0.0 %)
```

Berdasarkan hasil yang diperoleh, ditemukan beberapa kolom yang memiliki data pencilan, yakni kolom *temp*, *temp_min*, *temp_max*, *pressure*, *humidity*, *wind_speed* dengan proporsi yang terbilang kecil

1.4. *Imbalance dataset*

```
# Balance of data
df_true = data[data["raining"] == True]
df_false = data[data["raining"] == False]
print( "Class True = ", len(df_true), "; Class False =", len(df_false))
```

✓

```
Class True =  34901 ; Class False = 230023
```

Berdasarkan hasil yang diperoleh, ditemukan adanya *imbalance dataset*. Kelas *True* merupakan kelas minoritas dengan jumlah 34901 dan kelas *False* merupakan kelas mayoritas dengan jumlah 230023. Kelas mayoritas melebihi kelas minoritas 1 : 6,6.

2. Penanganan dari hasil analisis data

2.1. *Duplicate value*

Pengangan *duplicate value* dilakukan dengan menyisakan satu *row* dari kumpulan *rows* dengan nilai sama.

2.2. *Missing value*

Karena tidak terdapat *missing value,* tidak ada penanganannya.

2.3. *Outlier*

Tidak dilakukan mekanisme penghapusan atau imputasi data yang merupakan *outlier*.

2.4. *Imbalance dataset*

Imbalance dataset ditangani dengan teknik *oversampling*. Terdapat dua kelas yaitu kelas *True* dan kelas *False*. Kelas *True* merupakan kelas minoritas yang akan diduplikasi sehingga memiliki jumlah yang mirip dengan kelas *False*.

3. Justifikasi teknik-teknik yang dipilih

3.1. Duplicate value

Penanganan dengan menghapus *duplicate value* dilakukan untuk menghindari terjadinya *overfitting* pada model ketika dipakai untuk memprediksi data baru. Alasan lainnya adalah terkait efisiensi, dengan mengurangi ukuran data, performa dalam membangun model tentu akan memberikan efisiensi yang relatif lebih baik.

3.2. Missing value

Tidak terdapat *missing value*.

3.3. Outlier

Berdasarkan *domain knowledge* yang kami miliki ketika melihat data *outlier* yang ada menggunakan visualisasi *boxplot*, hampir semua *instance* yang merupakan *outlier* memiliki angka dalam *range* yang masuk akal dan kami melihat *outlier* yang ada dapat memberikan informasi yang bermanfaat.

3.4. Imbalance dataset

Penanganan *imbalance dataset* dengan teknik *oversampling* dilakukan untuk menyamakan jumlah data kedua kelas agar tidak terjadi bias terhadap kelas mayoritas. Pemilihan teknik *oversampling* daripada *undersampling* adalah agar informasi pada kelas mayoritas tidak hilang.

4. Perubahan yang dilakukan pada jawaban poin 1—5

Pada bagian strategi eksperimen, kami menghapus strategi *feature selection*.

5. Desain eksperimen

5.1. Tujuan eksperimen

Eksperimen dilakukan untuk mengoptimalkan performa model dengan menggunakan data yang ada. Performa yang diinginkan dari eksperimen ini adalah seberapa benarnya prediksi model pada kelas *True* karena merupakan kelas minoritas dari *imbalance dataset*. Performa akan dicari adalah nilai f1 dan *recall* yang besar tanpa mengorbankan *precision*.

5.2. Variabel dependen dan independen

Pada kasus ini, variabel independennya adalah fitur-fitur yang ada, yakni *hour*, *temp*, *temp_min*, *temp_max*, *pressure*, *humidity*, *wind_speed*, dan *wind_deg*. Sedangkan variabel dependennya adalah kolom target, yakni *raining*

5.3. Strategi eksperimen

Eksperimen dilakukan *hyperparameter tuning* pada parameter *logistic regression*.

5.4. Skema Validasi

Validasi dilakukan dengan menggunakan data validasi (df_val). Data validasi akan diprediksi kelasnya dengan model. Hasil prediksi tersebut akan dibandingkan dengan kelas aslinya. Metrik yang digunakan adalah *accuracy*, *precision*, *recall*, dan f1. Model yang dipilih adalah model yang memiliki nilai *recall* dan f1 yang baik tanpa mengorbankan *precision*.

6. Hasil eksperimen

Model	***Accuracy***	***Precision***	***Recall***	***f1***
Baseline	0.8728	0.5819	0.128	0.2099
Eksperimen	0.7177	0.2912	0.7468	0,419

7. Analisis dari hasil eksperimen

Model *baseline* memiliki akurasi yang tinggi, yakni 0.8728, yang berarti model dapat memprediksi 87% dari keseluruhan *test set* dengan benar. Namun, ketika memprediksi *true positive* (kasus *raining* = *true* atau *raining* terjadi), model memberikan banyak prediksi keliru. Nilai *precision* sebesar 0.5819 mengindikasikan bahwa dari seluruh hasil prediksi model yang memprediksi *raining* = *true*, hanya 58% yang benar. Nilai *f1* cukup kecil yang menunjukkan ketidakseimbangan antara nilai *precision* dan *recall*.

Setelah dilakukan eksperimen, nilai akurasi cukup tinggi, yakni dapat memprediksi 71% *test set* dengan benar. Ketika memprediksi ketika memprediksi *true positive* (kasus *raining* = *true* atau *raining* benar terjadi), model memberikan prediksi yang jauh lebih baik dibandingkan sebelum dilakukan eksperimen. Namun, ketika model memprediksi *raining* = *true*, hanya 29% yang prediksinya tepat. Walaupun nilai *f1* yang menunjukkan keseimbangan *trade-off* antara nilai *precision* dan *recall* lebih baik dibandingkan model *baseline*, skor ini tergolong kecil.

8. Kesimpulan

Berdasarkan metrik evaluasi yang diperoleh, terdapat peningkatan pada metrik *recall* dan *f1*, dan sebaliknya terdapat pengurangan angka *accuracy* dan *precision* setelah dilakukan eksperimen. Dengan demikian, dengan memperhatikan *trade-off* antara metrik yang ada, model hasil eksperimen memberikan hasil yang lebih baik karena untuk kasus memprediksi apakah hujan benar terjadi atau tidak, kita lebih memperhatikan *cost* dari *false negative* (nilai *recall*), yakni rain sebenarnya bernilai *true* tetapi prediksinya *false*, seperti tidak membawa payung ketika sebenaranya hujan, akan jauh lebih diperhatikan.

Pembagian tugas/kerja per anggota kelompok

NIM	Nama	Bagian pengerjaan
13520138	Gerald Abraham Sianturi	Baseline, analisis duplicate value, analisis missing value, analisis outlier, penanganan duplicate value, penanganan missing value, penanganan outlier, variabel dependen dan independen, strategi eksperimen. Implementasi penanganan duplicate value, implementasi penanganan missing value, implementasi penanganan outlier, hyperparameter tuning, validasi.

		Laporan
13520162	Daffa Romyz Aufa	Analisis imbalance dataset, penanganan imbalance dataset, teknik encoding, tujuan eksperimen, skema validasi. Implementasi penanganan imbalance dataset. Laporan