Praktikum

IF3270 Pembelajaran Mesin

Anggota Kelompok:

Vionie Novencia	13520006
Jaya Mangalo	13520015

Program Studi Teknik Informatika

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

2021

1. Analisis Data

1.1 Hasil Analisis Data

Pada dataset Denpasar Weather Data, terdapat 7253 data duplikat, 0 data hilang, 34901 data yang menyatakan raining, dan 230023 data yang menyatakan tidak raining. Sedangkan untuk *outliers*, terdapat 1458 *outliers* pada kolom temp, 1716 *outliers* pada kolom temp_min, 547 *outliers* pada kolom temp_max, 1067 *outliers* pada kolom pressure, 231 *outlier*s pada kolom humidity, 3439 *outliers* pada kolom wind_speed, dan tidak terdapat *outliers* pada kolom hour dan wind_deg.

1.2 Penanganan Hasil Analisis Data

1.2.1. Penanganan data duplikat

Data yang duplikat ditangani dengan melakukan penghapusan. Penghapusan ini dilakukan dengan menggunakan fungsi drop_duplicates(keep='first') dari library pandas.

1.2.2. Penanganan data yang hilang

Pada dataset Denpasar Weather Data, tidak terdapat data yang hilang, sehingga tidak dilakukan handling terhadap data yang hilang.

1.2.3. Penanganan *outliers*

Penanganan *outliers* dilakukan dengan mengganti semua nilai yang merupakan *outliers* dengan nilai median pada kolom yang sama.

1.2.4. Penanganan *imbalanced data*

Penanganan imbalanced data menggunakan library imblearn yang memiliki implementasi SMOTE dan *RandomUnderSampling*.

1.3 Justifikasi teknik-teknik yang dipilih

Pada penanganan data duplikasi, data yang duplikat dihapus karena apabila data tersebut dibiarkan, dapat menyebabkan bias. Sedangkan pada *outliers*, data yang merupakan *outliers* diganti dengan median pada kolom yang sama. Hal ini dilakukan agar menghindari adanya noise yang disebabkan oleh data tersebut. SMOTE dipilih karena data tidak memiliki categorical data

sehingga tidak memerlukan SMOTENC, sementara random undersampling dipilih karena cepat dan efisien untuk memberi perbandingan terhadap SMOTE yang lebih complex

2. Eksperimen

2.1 Desain Eksperimen

2.1.1 Tujuan Eksperimen

Tujuan dari eksperimen ini adalah untuk mencoba melakukan improvement terhadap baseline model yang sudah dibuat agar dapat lebih robust dalam melakukan prediksi data tidak terlihat.

5.2 Variabel Dependen dan Independen

Variabel dependen adalah fitur/kolom yang ingin diprediksi dan variabel independen adalah fitur/kolom yang digunakan untuk melakukan prediksi variabel dependen.

Untuk dataset ini, Variabel dependen merupakan kolom ***raining*** dan Variabel independen adalah kolom hour, temp, temp_min, temp_max, pressure, humidity, wind_speed, wind_deg.

5.3 Strategi Eksperimen

- Data preprocessing: Melakukan cleaning dan modifikasi data seperti handling missing values, duplicate data, outliers, imbalanced data.
- Feature selection: Tidak diperlukan karena semua data sudah cukup relevan.
- Hyperparameter tuning: Melakukan tuning parameter dengan menggunakan teknik grid search.

5.4 Skema Validasi

Validasi menggunakan k-fold *cross-validation* menggunakan data training yang sudah dipisah sebelumnya.

2.2 Hasil Eksperimen

2.2.1 Baseline Model

```
Accuracy: 0.8718763074816477
Precision: 0.581924577373212
Recall: 0.128058377450279
F1: 0.20992142605840272

array([[44950,   643],
       [ 6094,   895]], dtype=int64)
```

2.2.2 Oversampled Data(without nulls, duplicates, and outliers)

```
Accuracy: 0.7142748469057852
Precision: 0.2842257908588002
Recall: 0.7571898697953927
F1: 0.4133083411433927

array([[32266, 13327],
       [ 1697,  5292]], dtype=int64)
```

2.2.3 Undersampled Data(without nulls, duplicates, and outliers)

```
Accuracy: 0.7164809250313796
Precision: 0.28519503065154883
Recall: 0.752182000286164
F1: 0.4135787900243883

array([[32417, 13176],
       [ 1732,  5257]], dtype=int64)
```

2.2.4 Oversampled data (without nulls, duplicates, and outliers) + Hyperparameter Tuning and Cross Validation

```
Accuracy: 0.7166140504355103
Precision: 0.28516346258281744
Recall: 0.7513235083702962
F1: 0.4134157382986262

array([[32430, 13163],
       [ 1738,  5251]], dtype=int64)
```

2.2.5 Data Prediksi

Berikut merupakan deskripsi data yang terprediksi Rain = True oleh model Oversampling with Hyperparameter Tuning.

	hour	temp	temp_min	temp_max	pressure	humidity	wind_speed	wind_deg
count	18414.000000	18414.000000	18414.000000	18414.000000	18414.000000	18414.000000	18414.000000	18414.000000
mean	13.854459	26.019186	25.907506	26.127169	1009.869165	88.899479	3.587077	187.483491
std	6.978285	1.304980	1.323982	1.338282	2.267352	4.531352	2.080545	93.121207
min	0.000000	14.850000	7.000000	19.000000	1000.000000	69.000000	0.020000	0.000000
25%	9.000000	25.120000	25.000000	25.200000	1008.100000	86.000000	2.100000	110.000000
50%	16.000000	26.000000	26.000000	26.000000	1010.000000	88.000000	3.100000	151.000000
75%	20.000000	26.950000	26.900000	27.000000	1011.200000	93.000000	4.860000	272.000000
max	23.000000	31.700000	31.000000	36.200000	1018.000000	100.000000	31.900000	360.000000

```
Hujan average terjadi pada jam:                13.854458564135983, dengan kisaran 0 - 23
Hujan average terjadi pada temperature:        26.019185945476266, dengan kisaran 14.85 - 31.7
Hujan average terjadi pada pressure:           1009.869164765939, dengan kisaran 1000.0 - 1018.0
Hujan average terjadi pada humidity:           88.89947865754317, dengan kisaran 69 - 100
Hujan average terjadi pada kecepatan angin:    3.5870772238514177, dengan kisaran 0.02 - 31.9
Hujan average terjadi pada arah angin:         187.4834908222005, dengan kisaran 0 - 360
```

2.3 Analisis Eksperimen

Hasil Baseline memiliki akurasi tinggi(0.87) tetapi memiliki F1 score yang sangat rendah(0.21).

Metode desain eksperimen yang paling efektif merupakan melakukan pembersihan data terutama pada bagian oversampling dan undersampling. Model yang di-train setelah pembersihan data tersebut memiliki penuruan akurasi ke (0.71), tetapi memiliki F1 score yang lebih tinggi(0.41).

Hyperparameter Tuning dan Cross Validation tidak memiliki impact yang terlalu bermakna, hasil setelah tahap tersebut memiliki hasil yang kurang lebih sama, akurasi 0.71 dan F1 score 0.41.

Dari data metrik model tersebut, dapat disimpulkan bahwa model baselina dapat memprediksi kelas mayoritas dengan baik (Rain = False), tetapi sangat jelek dalam memprediksi kelas minoritas (Rain = True). Hal ini dibuktikan dalam nilai akurasi yang tinggi tetapi nilai precision yang rendah.

Model yang telah diberlakukan oversampling dan undersampling sedikit lebih baik, walaupun nilai akurasi turun, ia lebih baik dalam memprediksi kelas minoritas yaitu kelas yang lebih penting dalam dataset ini.

3. Kesimpulan

3.1 Pembelajaran mesin dapat digunakan untuk mengolah dan analisis data untuk melakukan prediksi sebuah data lain. Dengan mengetahui data cuaca denpasar, kita dapat melakukan prediksi apakah akan hujan atau tidak.

3.2 Pengembangan model memerlukan metode dan teknik lain seperti penangan data serta penentuan algoritma untuk meningkatkan kualitas algoritma atau model yang dibuat. Sebuah contoh penanganan yang efektif pada praktikum ini adalah melakukan *handling imbalanced dataset* yang menaikkan performa F1 Score.

3.3 Hasil metrik dan data dari beberapa model dan metode yang berbeda dapat digunakan untuk memberi insight tentang data. Contohnya dapat mendapatkan insight data seperti biasanya terjadi pada suhu 26C.

4. Pembagian Tugas

13520006	Vionie Novencia	2,3,4,6, Laporan
13520015	Jaya Mangalo	1,5,6,7, Laporan