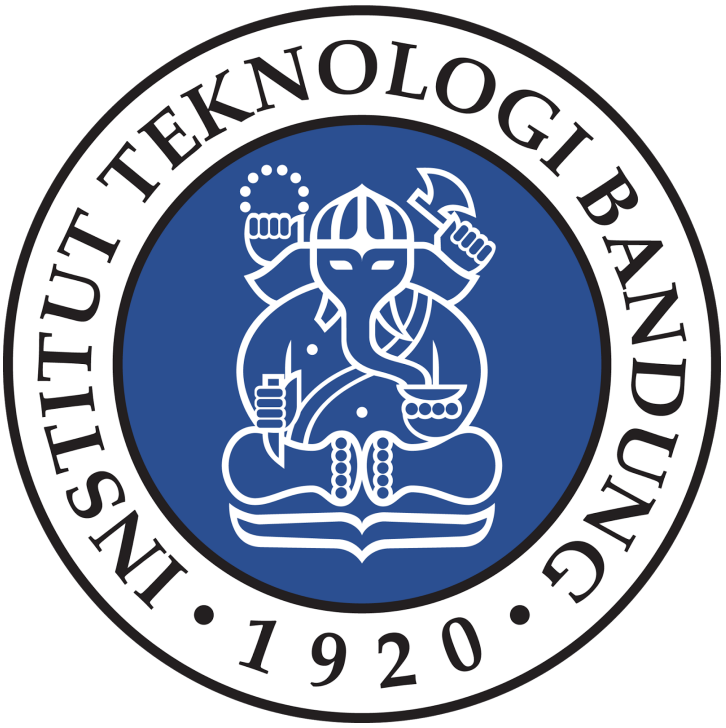
IF3270 - Pembelajaran Mesin

Praktikum



Disusun oleh:

13520133 - Jevant Jedidia Augustine

13520160 - Willy Wilsen

Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung
2023

Dataset

Dataset yang digunakan adalah [openweatherdata-denpasar-1990-2020v0.1-simplified.csv]

Hasil Analisis Data

- ## Duplicate value

  Terdapat 7253 data duplikat pada dataset keseluruhan yang diberikan.

- ## Missing value

```
hour          0
temp          0
temp_min      0
temp_max      0
pressure      0
humidity      0
wind_speed    0
wind_deg      0
raining       0
```

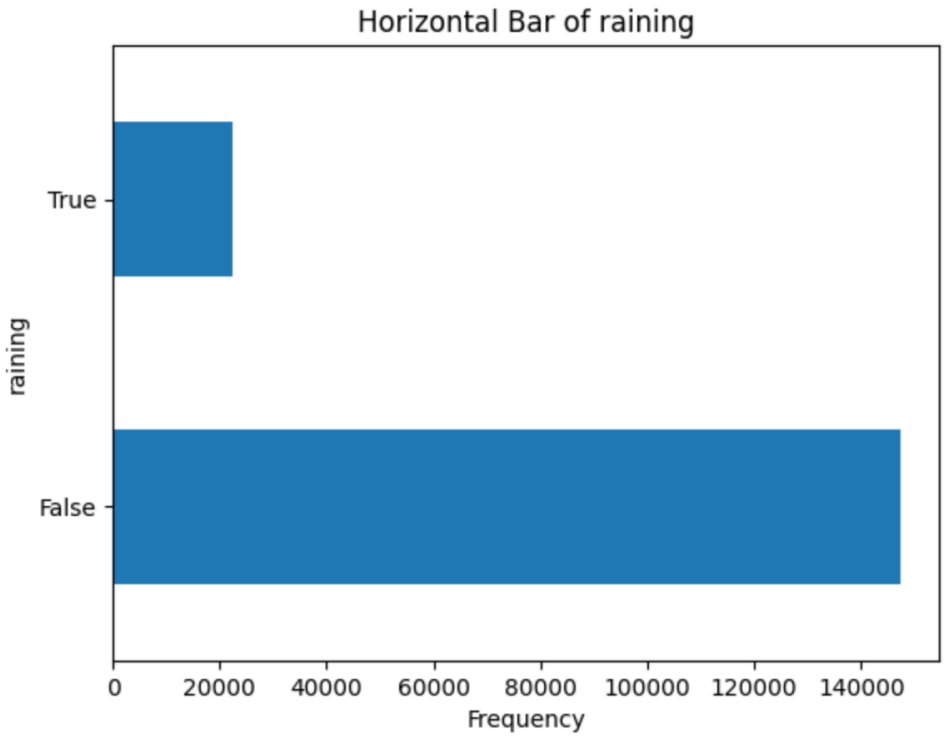
  Tidak terdapat *missing values* pada dataset keseluruhan yang diberikan.

- ## Outlier

```
hour             0
temp          1458
temp_min      1716
temp_max       547
pressure      1067
humidity       231
wind_speed    3439
wind_deg         0
```

  Fitur *hour* dan *wind_deg* tidak memiliki outlier tetapi fitur lainnya memiliki outlier yang dapat dibilang relatif kecil bila dibandingkan dengan ukuran data keseluruhan yaitu fitur *temp* dengan jumlah 1458 outlier, fitur *temp_min* dengan jumlah 1716 outlier, fitur *temp_max* dengan jumlah 547 outlier, fitur *pressure* dengan jumlah 1067 outlier, fitur *humidity* dengan jumlah 231 outlier, dan fitur *wind_speed* dengan jumlah 3439 outlier.

- Balance of data



Terdapat ketidakseimbangan pada data latih. Label False pada raining sangat besar apabila dibandingkan dengan label True.

Penanganan dari Hasil Analisis Data

- Duplicate value

  Hal yang dilakukan untuk menangani data yang duplikat adalah dengan membuang data duplikat tersebut.

- Missing value

  Tidak terdapat *missing values* pada data yang diberikan, maka tidak perlu ditangani.

- Outlier

  Karena jumlah outlier relatif kecil terhadap jumlah dataset, maka tidak perlu ditangani.

- Balance of data

  Hal yang dilakukan untuk menangani *imbalanced data* adalah dengan melakukan oversampling atau undersampling pada dataset.

Justifikasi Teknik-teknik yang Dipilih

1. Membuang data yang duplikat
2. Melakukan oversampling atau undersampling pada imbalanced dataset
3. Melakukan encoding pada target dataset

Perubahan yang Dilakukan pada Poin 5

Strategi eksperimen yang diterapkan sebelum perubahan adalah:

1. Membuang data yang duplikat
2. Dilakukan oversampling untuk mengatasi imbalanced data
3. Melakukan encoding pada target
4. Membagi data menjadi data latih, data validasi, dan data test
5. Menggunakan Logistic Regression untuk membuat model dari data latih
6. Menggunakan hyperparameter tuning dengan Grid Search untuk menemukan parameter terbaik dari Logistic Regression
7. Melakukan skema validasi terhadap model yang dibuat
8. Menghitung nilai metrik dan confusion matrix dari model

Strategi eksperimen yang diterapkan setelah perubahan adalah:

1. Membuang data yang duplikat
2. Dilakukan oversampling untuk mengatasi imbalanced data
3. Melakukan encoding pada target
4. Membagi data menjadi data latih, data validasi, dan data test
5. Menggunakan Logistic Regression untuk membuat model dari data latih
6. Menghitung nilai metrik dan confusion matrix dari model baseline Logistic Regression
7. Menggunakan hyperparameter tuning dengan Grid Search untuk menemukan parameter terbaik dari Logistic Regression
8. Melakukan skema validasi terhadap model yang dibuat
9. Menghitung nilai metrik dan confusion matrix dari model Logistic Regression dengan parameter yang sudah didapat dari hyperparameter tuning

Alasan dilakukannya perubahan adalah untuk menghitung nilai metrik dan confusion matrix pada model baseline Logistic Regression sehingga dapat dibandingkan dengan model Logistic Regression dengan parameter yang sudah didapat dari hyperparameter tuning.

Desain Eksperimen

- ## Tujuan eksperimen

  Tujuan dari eksperimen adalah untuk membuat sebuah model yang dapat memprediksi apakah akan hujan berdasarkan beberapa variabel masukan.

- ## Variabel dependen dan independen

  Variabel dependen adalah 'hour', 'temp', 'temp_min', 'temp_max', 'pressure', 'humidity', 'wind_speed', dan 'wind_deg'. Variabel independen adalah 'raining'.

- ## Strategi eksperimen

  Strategi eksperimen yang diterapkan adalah:

1. Membuang data yang duplikat
2. Dilakukan oversampling untuk mengatasi imbalanced data
3. Melakukan encoding pada target
4. Membagi data menjadi data latih, data validasi, dan data test
5. Menggunakan Logistic Regression untuk membuat model dari data latih
6. Menghitung nilai metrik dan confusion matrix dari model baseline Logistic Regression
7. Menggunakan hyperparameter tuning dengan Grid Search untuk menemukan parameter terbaik dari Logistic Regression
8. Melakukan skema validasi terhadap model yang dibuat
9. Menghitung nilai metrik dan confusion matrix dari model Logistic Regression dengan parameter yang sudah didapat dari hyperparameter tuning

- ## Skema validasi

  Skema validasi yang digunakan adalah K-fold cross-validation

Hasil Eksperimen

Pada model baseline Logistic Regression, didapatkan 26584 data berlabel True Positive, 8964 data berlabel False Negative, 10814 data berlabel False Positive, dan 24931 data berlabel True Negative dari keseluruhan data. Accuracy, Precision, Recall, dan F1 yang didapatkan berturut-turut bernilai 72.25%, 71.08%, 74.78%, dan 72.88%.

```
Accuracy: 0.7225814596103404
Precision: 0.7108401518797797
Recall: 0.7478339147068752
F1: 0.728867929701423
Confusion Matrix:
```

```
array([[24931, 10814],
       [ 8964, 26584]])
```

Kemudian, dilakukan hyperparameter tuning untuk menentukan parameter terbaik pada model Logistic Regression. Param terbaik yang didapatkan adalah tol bernilai 0.0001 dan C bernilai 10000.

```
Best params:
tol: 0.0001
C: 10000
```

Lalu, pada model Logistic Regression dengan parameter yang sudah didapat dari hyperparameter tuning, didapatkan 26563 data berlabel True Positive, 8985 data berlabel False Negative, 10815 data berlabel False Positive, dan 24930 data berlabel True Negative dari keseluruhan data. Accuracy, Precision, Recall, dan F1 yang didapatkan berturut-turut bernilai 72.27%, 71.06%, 74.72%, dan 72.84%.

```
Accuracy: 0.7222728739146901
Precision: 0.7106586762266573
Recall: 0.7472431641723867
F1: 0.7284918958944684
Confusion Matrix:
```

```
array([[24930, 10815],
       [ 8985, 26563]])
```

Analisis dari Hasil Eksperimen

Apabila hasil baseline dibandingkan dengan hasil setelah dilakukan hyperparameter tuning, tidak terlihat perubahan yang signifikan sehingga hyperparameter tuning pada model ini tidak terlalu meningkatkan performa pada model.

Kesimpulan

Berdasarkan hasil prediksi yang dihasilkan, karakteristik kondisi hujan adalah sebagai berikut.

1. Tidak dipengaruhi oleh ‘hour’.
2. Saat ‘temp’ < 20, maka kemungkinan besar terjadi hujan dan saat ‘temp’ > 32.5, maka kemungkinan besar tidak terjadi hujan.
3. Saat ‘temp_min’ < 18, maka kemungkinan besar terjadi hujan dan saat ‘temp_min’ > 32, maka kemungkinan besar tidak terjadi hujan.
4. Saat ‘temp_max’ < 20, maka kemungkinan besar terjadi hujan dan saat ‘temp_max’ > 32, maka kemungkinan besar tidak terjadi hujan.
5. Tidak dipengaruhi oleh ‘pressure’
6. Saat ‘humidity’ < 70, maka kemungkinan besar tidak terjadi hujan.
7. Tidak dipengaruhi oleh ‘wind_speed’
8. Tidak dipengaruhi oleh ‘wind_deg’

Pembagian Tugas/Kerja per Anggota Kelompok

NIM - Nama	Pembagian Tugas
13520133 - Jevant Jedidia Augustine	1. Membuat baseline dengan menggunakan model logistic regression. 2. Analisis data. 3. Rencana penanganan dari hasil analisis data. 4. Teknik encoding yang digunakan terhadap data yang disediakan 5. Desain eksperimen 6. Implementasi strategi eksperimen dan skema validasi 7. Kesimpulan analisis karakteristik kondisi hujan. 8. Laporan
13520160 - Willy Wilsen	1. Membuat baseline dengan menggunakan model logistic regression. 2. Analisis data. 3. Rencana penanganan dari hasil analisis data. 4. Teknik encoding yang digunakan terhadap data yang disediakan 5. Desain eksperimen 6. Implementasi strategi eksperimen dan skema validasi 7. Kesimpulan analisis karakteristik kondisi hujan. 8. Laporan