# PRAKTIKUM IF3270 PEMBELAJARAN MESIN

**Kelas 03**

**Semester II Tahun Akademik 2023/2024**


**Disusun Oleh:**

**Anggota Kelompok:**

| | |
|---|---|
| **Azmi Hasna Zahrani** | **13521006** |
| **Muhamad Salman Hakim Alfarisi** | **13521010** |



**INSTITUT TEKNOLOGI BANDUNG**

**BANDUNG**

**2023**

# I. Analisis Data

## A. Hasil Analisis Data

Pada dataset Diabetes, terdapat 2329 *duplicate value,* 0 *missing value*, 11 kolom data *outlier,* serta adanya *imbalanced dataset* dengan kelas *True* sebagai kelas minoritas dengan total sebanyak 6946 dan kelas *False* yang merupakan kelas mayoritas dengan total sebanyak 43790. Data *outlier* ditemukan pada kolom BMI sebanyak 1979, Stroke sebanyak 2043, HeartDeseaseorAttack sebanyak 4693, PhysActivity sebanyak 12455, Veggies sebanyak 9494, HvyAlcoholConsump sebanyak 2774, AnyHealthcare sebanyak2402, GenHlth sebanyak 2365, MenHlth sebanyak 7308, PhysHlth sebanyak 8198, dan DiffWalk sebanyak 8588.

## B. Penanganan Hasil Analisis Data

1. Penanganan *Duplicate Value*
   *Duplicate value* dapat ditangani dengan melakukan penghapusan pada data duplikat.
2. Penanganan *Missing Value*
   *Missing value* dapat ditangani dengan menggantikan missing value dengan median. Akan tetapi pada data Diabetes tidak ditemukan missing value.
3. Penanganan *Outliers*
   *Outlier* dapat ditangani dengan melakukan penghapusan pada data *outlier* dan mengisi dengan median.
4. Penanganan *Imbalanced Data*
   *Imbalanced data* dapat ditangani dengan *oversampling* agar data pada kelas mayoritas tidak hilang.

## C. Justifikasi Teknik-Teknik yang Dipilih

1. Penanganan *Duplicate Value*
   Penanganan *duplicate value* dengan penghapusan data duplikat dilakukan untuk menghindari *overfitting* pada model saat akan dilakukan prediksi data baru.
2. Penanganan *Missing Value*
   Tidak dilakukan penanganan pada *missing value* karena tidak ada *missing value.*
3. Penanganan *Outliers*
   Penanganan *outliers* dengan menghapus data *outliers* dan mengisinya dengan median dilakukan untuk menghindari data noise yang disebabkan oleh *outliers* tersebut.
4. Penanganan *Imbalanced Data*

Penanganan *imbalanced data* dengan *oversampling* dilakukan untuk menyamakan jumlah kedua kelas agar tidak terjadi bias terhadap kelas mayoritas.

# II. Eksperimen

## A. Desain Eksperimen

1. Tujuan Eksperimen
   Tujuan dari eksperimen ini adalah untuk meningkatkan performa model baseline agar lebih meningkatkan robustness dalam prediksi data.
2. Variabel Dependen dan Independen
   Variabel dependen pada kasus ini adalah kolom target, yaitu kolom Diabetes. Sedangkan, variabel independen adalah fitur-fitur yang ada pada data yaitu HighBP, HighChol, BMI, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education, Income.
3. Strategi Eksperimen
   a) Data preprocessing: melakukan preprocessing data yaitu cleaning dan modifikasi data seperti encoding variabel kategorikal dan normalisasi data numerik.
   b) Hyperparameter tuning: melakukan tuning pada parameter *logistic regression*.
4. Skema Validasi
   Validasi dilakukan menggunakan data validasi (df_val) menggunakan k-fold cross-validation untuk memastikan keandalan hasil.

## B. Hasil Eksperimen

Hasil prediksi menggunakan model *logistic regression baseline* sebagai berikut:

- *Accuracy*: 0.8697955161369796
- *Precision*: 0.5570469798657718
- *Recall*: 0.152153987167736
- *F1 Score*: 0.2390208783297336

Hasil prediksi menggunakan model *logistic regression undersampling* setelah dilakukan *hyperparameter tuning* sebagai berikut:

- Accuracy: 0.7351564424735156
- Precision: 0.30352504638218925
- Recall: 0.7497708524289642
- F1 Score: 0.43211833069202327

Skor validasi menggunakan parameter *logistic regression* terbaik menggunakan *k-fold cross-validation* menghasilkan skor 0.7422145449869436.

### C. Analisis Hasil Eksperimen

Perbandingan Model *Logistic Regression Baseline* dan Setelah *Hyperparameter Tuning* adalah sebagai berikut.

- *Accuracy*: Setelah dilakukan *hyperparameter tuning*, akurasi model mengalami penurunan dari 0.8698 menjadi 0.7352. Hal ini menunjukkan bahwa model *baseline* lebih baik dalam mengklasifikasikan data.
- *F1 Score*: *F1 score* menunjukkan peningkatan dari 0.2390 menjadi 0.4321.

Hasil skor validasi menggunakan *k-fold cross-validation* dengan parameter tuning terbaik menunjukkan skor sebesar 0.7422. Hal ini menunjukkan bahwa model setelah tuning memiliki performa yang cukup stabil.

## III. Kesimpulan

Berdasarkan hasil prediksi, model *logistic regression* setelah dilakukan *hyperparameter tuning* memiliki performa yang lebih baik dibandingkan dengan model baseline dari segi *f1 score*. Selain itu, hasil validasi menggunakan *k-fold cross-validation* juga menunjukkan bahwa model *logistic regression* setelah dilakukan *hyperparameter tuning* memiliki skor yang cukup stabil.

## IV. Pembagian Tugas

| NIM | Nama | Pembagian Tugas |
|---|---|---|
| 13521006 | Azmi Hasna Zahrani | 2, 3, 4, 5, laporan |
| 13521010 | Muhamad Salman Hakim Alfarisi | 1, 6, 7, laporan |