Laporan Praktikum

Disusun Oleh:

13521116 Juan Christopher Santoso

13521135 Nicholas Liem

Dosen Pengampu : Fariska Zakhralativa Ruskanda, S.T., M.T.

IF3270 - Pembelajaran Mesin

Program Studi Teknik Informatika

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

2024

Daftar Isi

Hasil Analisis Data

Preprocessing

Berikut adalah data hasil preprocessing kami yakni sekedar statistik singkat tentang data yang akan dianalisis:

- Data Size: 50736 rows, 20 cols
- Columns (Features): ['HighBP', 'HighChol', 'BMI', 'Smoker', 'Stroke', 'HeartDiseaseorAttack', 'PhysActivity', 'Fruits', 'Veggies', 'HvyAlcoholConsump', 'AnyHealthcare', 'GenHlth', 'MentHlth', 'PhysHlth', 'DiffWalk', 'Sex', 'Age', 'Education', 'Income', 'Diabetes']

Categorical Grouping

Berikut adalah hasil pembagian antara data yang bersifat kategorikal dan non-kategorikal:

1. Categorical: ['HighBP', 'HighChol', 'Smoker', 'Stroke', 'HeartDiseaseorAttack', 'PhysActivity', 'Fruits', 'Veggies', 'HvyAlcoholConsump', 'AnyHealthcare', 'DiffWalk', 'Sex', 'Diabetes']
2. Non-Categorical: ['BMI', 'GenHlth', 'MentHlth', 'PhysHlth', 'Age', 'Education', 'Income']

Missing Values

Tidak ada missing values pada dataset yang diberikan.

Data Outliers

Ada beberapa data outlier yang ditemukan pada data menggunakan pendekatan IQR:

1. BMI:
    a. Lower: 4
    b. Upper: 1975
2. GenHlth:
    a. Lower: 0
    b. Upper: 2365
3. MentHlth:
    a. Lower: 0
    b. Upper: 7308
4. PhysHlth:

a. Lower: 0
b. Upper: 8198

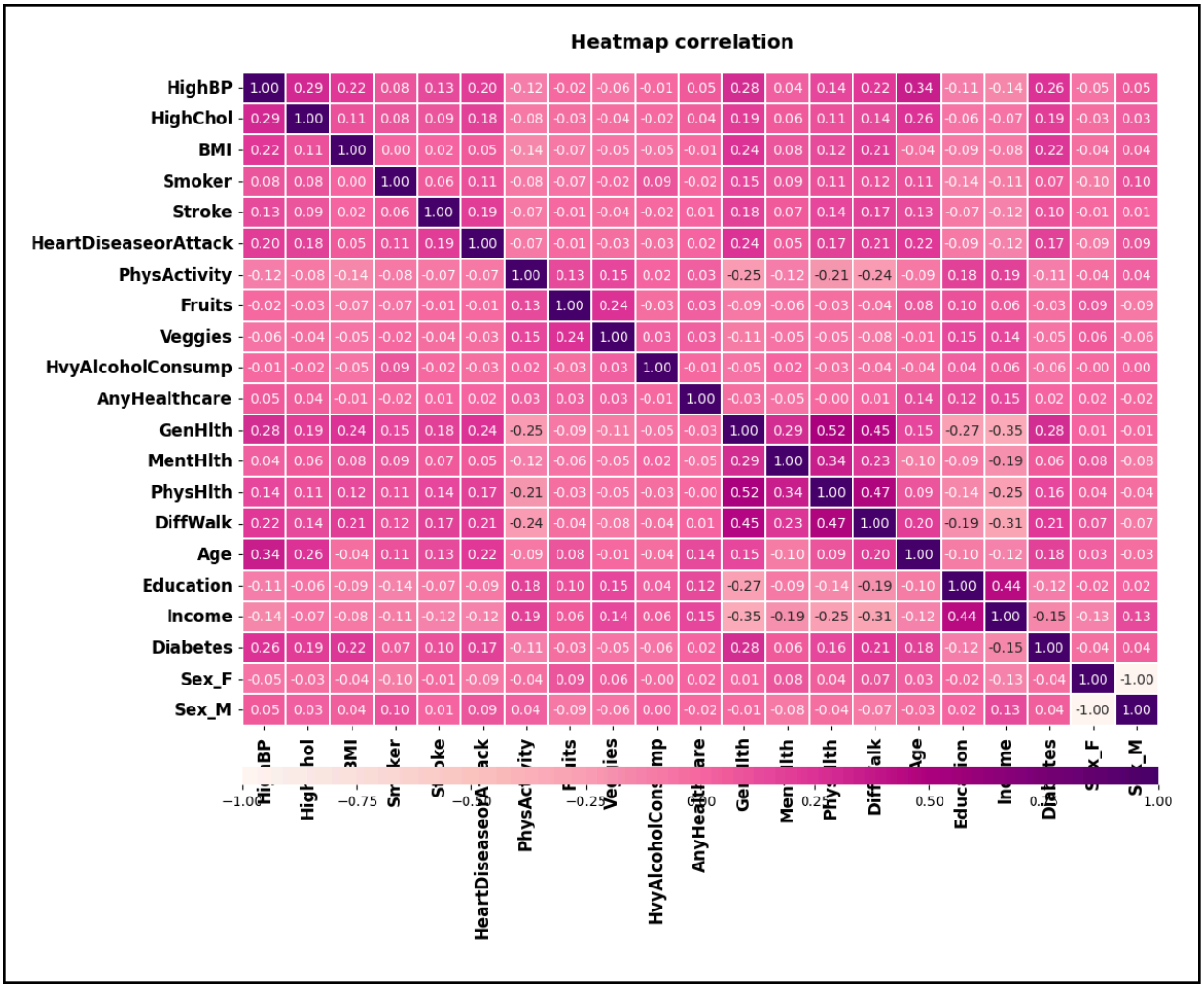
Duplicate Value

Dataset ini memiliki sebanyak 2329 duplicated values yang akan dibuang dan akan diambil yang unik saja.

Check Data Reasonableness

Untuk fitur BMI nilai yang tidak reasonable adalah BMI di atas 70 sehingga kemungkinan besar kami membuang data tersebut untuk memperbaiki data training nanti.

Heatmap Analysis

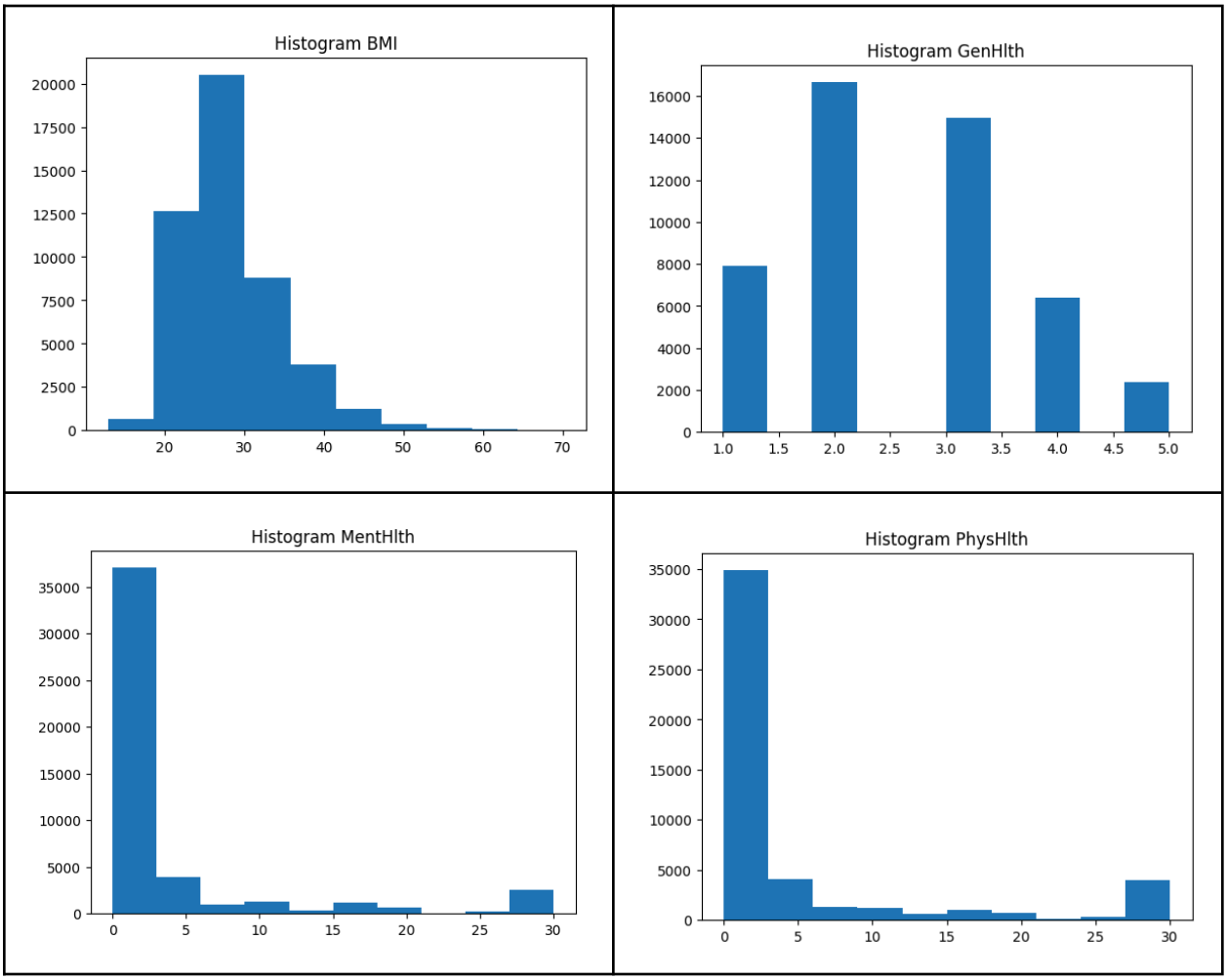


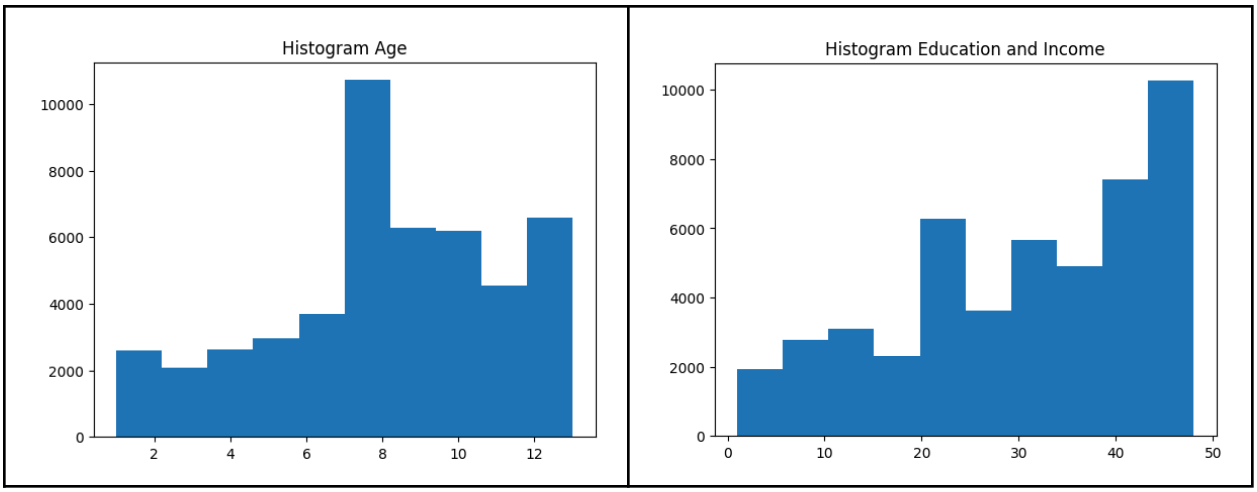
Berdasarkan hasil heatmap ini, kita dapat memeriksa beberapa hal.

1. Fitur Sex ternyata tidak terlalu begitu berkorelasi antar fitur lainnya sehingga kemungkinan besar fitur ini akan dibuang.
2. Fitur HvyAlcoholConsump juga tidak begitu berkorelasi dengan fitur lainnya sehingga kemungkinan besar fitur ini juga akan dibuang.
3. Beberapa fitur yang memiliki keterikatan yang tinggi adalah HighChol, BMI, Smoker, Stroke, HeartDiseaseorAttack, GenHlth, PhysHlth, MentHlth, DiffWalk, dan Age.
4. Fitur Education and Income kurang lebih hampir mirip jadi kemungkinan besar kedua fitur ini dapat digabungkan.

Histogram for Non-Categorical Data

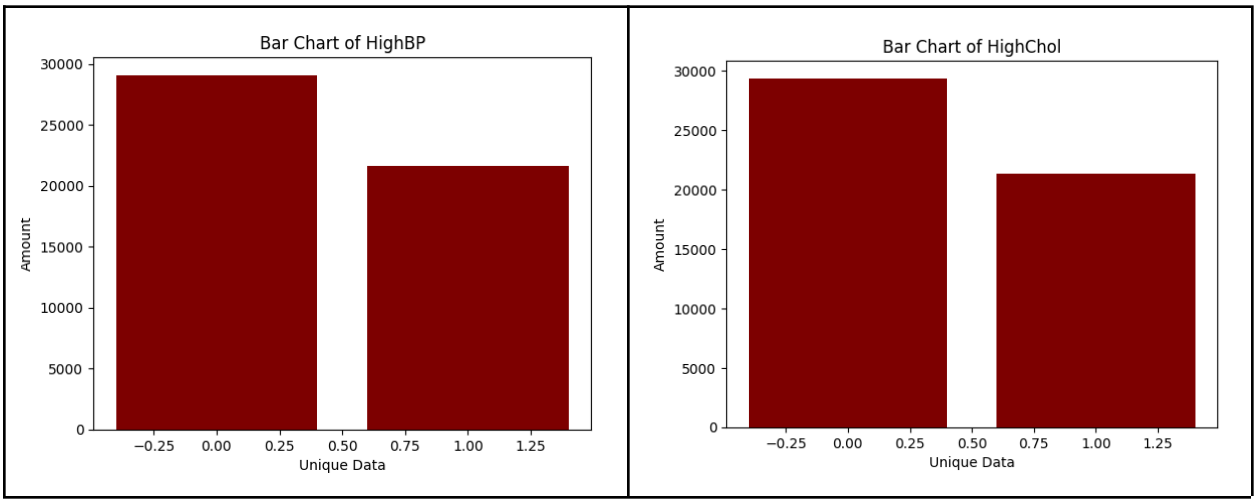
Histogram yang dibuat pada analisis data divisualisasikan hanya untuk data *non-categorical*. Data *non-categorical* umumnya bernilai numerik sehingga Histogram dapat menampilkan data tersebut dengan baik. Berikut adalah Histogram yang dibentuk.

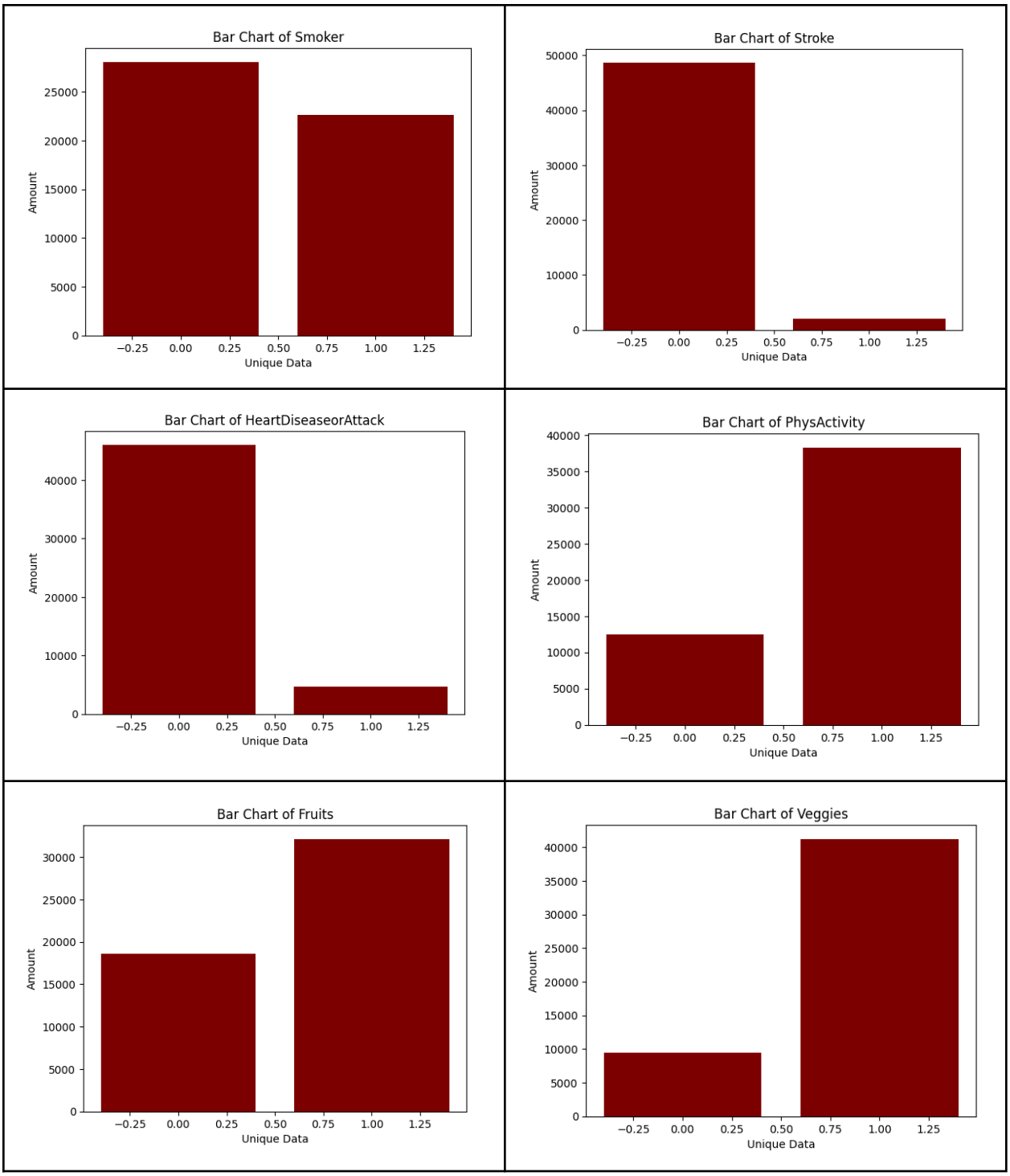


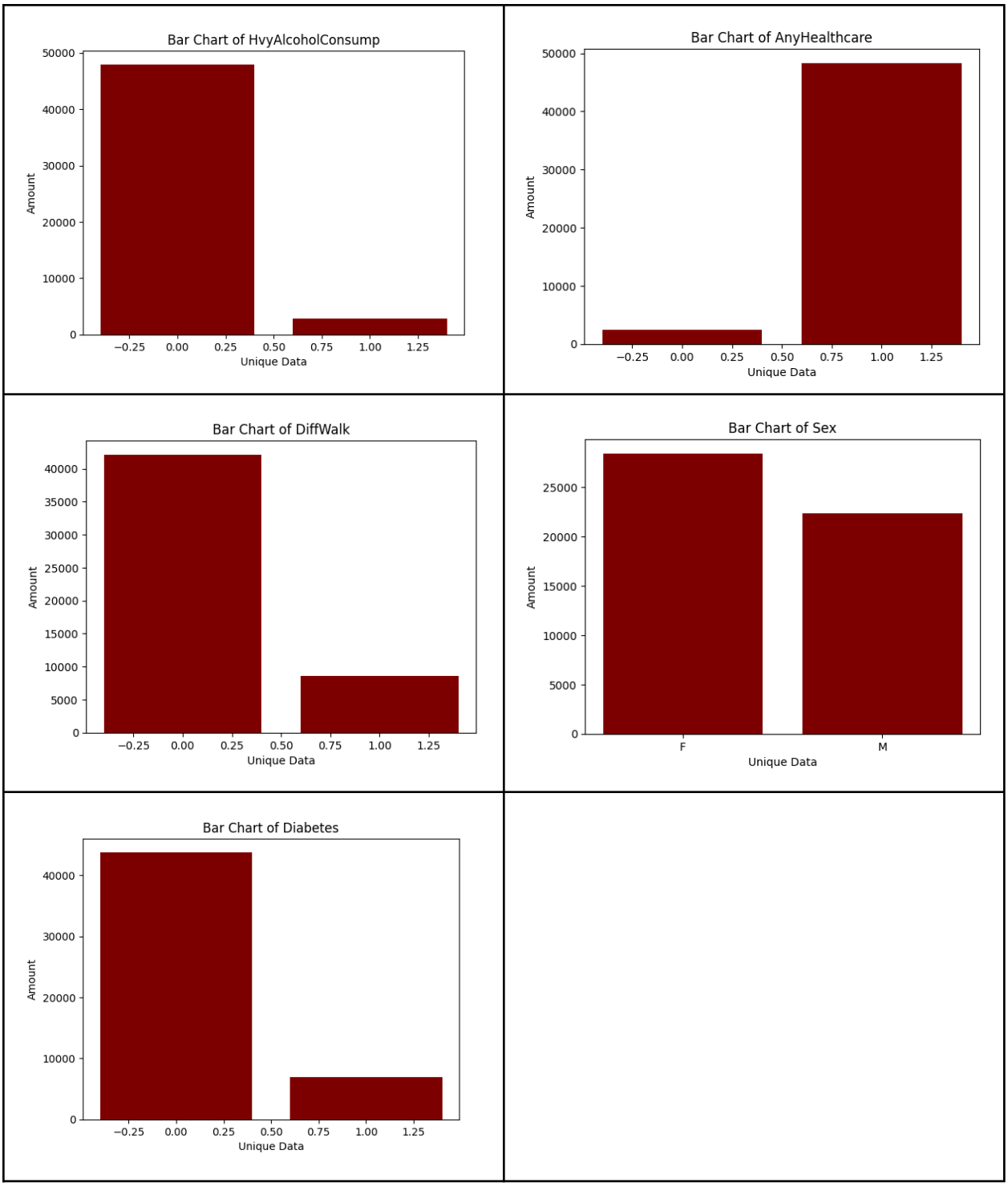


Bar Chart for Categorical Data

Bar Chart yang dibuat pada analisis data divisualisasikan hanya untuk data *categorical*. Data *categorical* umumnya memiliki variasi nilai unik yang sedikit. Dengan begitu, dapat dengan mudah digambarkan menggunakan Bar Chart. Berikut adalah Bar Chart yang dibentuk.




Bar Chart of Smoker
Amount
Unique Data
Bar Chart of Stroke
Amount
Unique Data
Bar Chart of HeartDiseaseorAttack
Amount
Unique Data
Bar Chart of PhysActivity
Amount
Unique Data
Bar Chart of Fruits
Amount
Unique Data
Bar Chart of Veggies
Amount
Unique Data


Bar Chart of HvyAlcoholConsump
Amount
Unique Data
Bar Chart of AnyHealthcare
Amount
Unique Data
Bar Chart of DiffWalk
Amount
Unique Data
Bar Chart of Sex
Amount
F
M
Unique Data
Bar Chart of Diabetes
Amount
Unique Data

Penanganan dari Hasil Analisis Data

Data Handling

Data handling terdiri dari dua prosedur yang dilakukan yakni *handling missing value*, *handle outliers and duplicated data*, dan *check data reasonableness*.

1. Handle missing value: Tidak ada missing value
2. Handle outliers and duplicated data:
    a. Buang semua yang outlier, tetapi setelah beberapa pertimbangan hanya fitur BMI saja yang akan dibuang karena setelah dibuang datanya menjadi terdistribusi normal, sebelumnya tidak.
    b. Untuk duplicated data, dibuang semua yang terduplikasi.
3. Check data reasonableness:
    a. Data BMI di atas 70 akan dibuang.

Encoding

Tipe encoding yang kami gunakan adalah *one-hot encoding* sehingga untuk fitur yang ter-encode akan ditambahkan kolom baru untuk menandakan atau pemisahan nilai-nilai pada fitur. Fitur yang kami *encode* adalah fitur Sex di mana jadinya ada dua kolom baru ‘Sex_M’ dan ‘Sex_F’.

Scaling

Kami menggunakan teknik scaling MinMax untuk data-data non kategorikal.

Drop Columns

Kolom yang kami drop karena kami pikir tidak begitu berpengaruh adalah ‘Sex_M’, ‘Sex_F’, dan ‘HvyAlcoholConsump’.

Feature Engineering

Kami menggabungkan fitur Education dan Income menjadi satu dengan mengalikan nilainya secara bersamaan dan menggantinya menjadi fitur ‘Education and Income’.

Justifikasi Teknik-Teknik yang Dipilih

Teknik Encoding

Teknik encoding yang kami gunakan adalah *one-hot encoding* karena tipe encoding ini dapat memperbaiki performa model dengan memberikan informasi tambahan tentang variabel kategorikal yang kami encode, misalnya dalam kasus ini adalah variabel Sex.

Teknik Scaling

Teknik encoding yang kami gunakan adalah *one-hot encoding* karena tipe encoding ini dapat memperbaiki performa model dengan memberikan informasi tambahan tentang variabel kategorikal yang kami encode, misalnya dalam kasus ini adalah variabel Sex.

Teknik Sampling

Teknik sampling yang kami gunakan adalah *oversampling* karena dataset yang diberikan itu tidak balance sehingga dibutuhkan teknik sampling yang supaya model dapat ditrain dengan baik. Alasan mengapa *oversampling* dipilih daripada *undersampling* adalah ketika kami melakukan eksperimen tersebut dan membandingkan kedua hasil berdasarkan metrik akurasi, f1, dan sebagainya untuk model yang diberikan teknik yang paling meningkatkan metrik adalah *oversampling*.

Teknik Skema Validasi

Skema validasi yang digunakan pada data train adalah menggunakan stratified 10-fold cross-validation karena skema ini biasanya dipilih untuk data yang imbalance. Mengapa k-fold cross-validation ini cocok?

1. Balanced splits: Standard K-Fold Cross-Validation memastikan bahwa setiap fold memiliki representasi distribusi kelasnya.
2. Better Generalization: Karena setiap fold memiliki distribusi yang mirip, model jadi tidak bias terhadap kelas tertentu.
3. Robust Performance Metrics: Dengan distribusi kelas yang seimbang pada setiap fold, performance metrics seperti accuracy, precision, dan recall menjadi lebih reliable.

Dengan memastikan apakah model tersebut sudah cukup baik atau tidak melalui nilai MSEnya yang dikumpulkan tiap fold dengan model akhir yang ingin diperiksa.

Referensi: https://www.linkedin.com/pulse/stratified-k-fold-cross-validation-in-depth-look-yeshwanth-n/

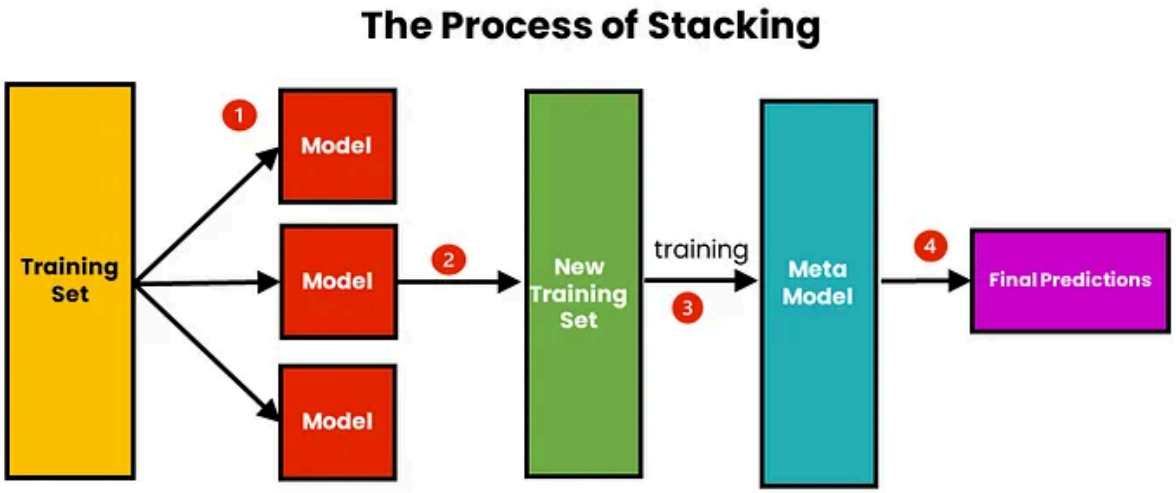
Teknik GridSearch

Teknik GridSearch dipilih untuk teknik *hyperparameter tuning* karena sifatnya yang mirip seperti cross-validation setiap parameter terbaik prediksi akan dilakukan. Selain itu, GridSearch juga melakukan evaluasi secara exhaustive terhadap setiap kombinasi, walaupun lama tetapi hasil *hyperparameter tuning* yang diberikan diharapkan menjadi yang terbaik.

Teknik Tuning untuk LogReg, KNN, dan MLP

Untuk tuning model Logistic Regression dan KNN kami menggunakan GridSearch sedangkan untuk MLP kami gunakan teknik *exhaustive loop* untuk menentukukan hidden layer yang terbaik bagi model yang akan dibangun.

Teknik Model Stacking



Teknik Model Stacking dipilih untuk mereduksi error yang disebabkan oleh generalisasi dari prediksinya dengan menggabungkan beberapa model karena lebih diverse dan independent.

Perubahan yang Dilakukan pada Jawaban Poin 1 - 5

1. Perubahan teks pada penjelasan #3 Summary adalah “Selain itu beberapa pemrosesan yang akan kami lakukan adalah pembuangan fitur **Sex, dan HvyAlcoholConsump** serta melakukan scaling.”
2. Menghapus bagian *adjustment* fitur target karena dirasa tidak penting.
3. Pada pelaksanaan *model stacking*, model KNN tidak jadi dipakai. Maka dari itu, *models* yang dipakai hanyalah Logistic Regression dan MLP.

Desain Eksperimen

Tujuan Eksperimen

Tujuan dari dilakukannya eksperimen ini adalah untuk memprediksi apakah seseorang memiliki diabetes atau tidak berdasarkan data-data yang diberikan. Berdasarkan data-data yang didapatkan kita mengaplikasikan dan melakukan tuning terhadap model-model klasifikasi yang diberikan dengan baseline model regresi logistik. Hasil dari tuning dan model terbaik akan dipilih sebagai prediktor terbaik untuk menentukan apakah seseorang diabetes atau tidak.

Variabel Dependen dan Independen

Variabel dependen pada eksperimen ini adalah fitur diabetes dan fitur-fitur lain selain diabetes adalah variabel-variabel independen.

Strategi Eksperimen

Dalam eksperimen kali ini, strategi yang kami gunakan mencakup penggantian model klasifikasi, pengaturan hyperparameter, model stacking, grid search, oversampling, dan undersampling. Alur yang digunakan adalah sebagai berikut:

1. Untuk data yang telah di preprocess, akan dilakukan splitting atau pemisahan. Tentunya pemisahan yang dilakukan adalah pemisahan data menjadi data train, valid, dan test. Penjelasan lebih lengkap dijelaskan pada bagian Skema Validasi.
2. Mengingat variabel dependen yang imbalance, akan dilakukan oversampling atau undersampling. Hal ini dilakukan untuk menghindari hasil yang bias. Namun, hal ini dapat berubah menyesuaikan apabila hasil menggunakan metode ini tidak memenuhi ekspektasi.
3. Model yang digunakan dapat berubah. Beberapa model yang akan kami gunakan adalah MLP, KNN, dan Logistic Regression.
4. Untuk setiap proses menggunakan masing-masing model, kami akan melakukan hyperparameter tuning. Salah satu yang akan kami gunakan adalah dengan grid search.
5. Mengingat terdapat 3 model yang kami gunakan, kami akan melakukan Model Stacking terhadap hasil prediksi dari seluruh model tersebut. Hal ini dilakukan guna menghasilkan prediksi yang lebih baik.

Skema Validasi

Skema validasi yang digunakan pada DoE data train adalah menggunakan stratified 10-fold cross-validation karena skema ini biasanya dipilih untuk data yang imbalance. Jadi awalnya data akan displit 80/20, 80 untuk train dan 20 untuk test, kemudian data train ini akan dilakukan skema validasi dengan 10-fold cross-validation untuk pelatihan model dengan memastikan apakah model tersebut sudah cukup baik atau tidak melalui nilai MSEnya. Validasi hasil akhir tetap menggunakan nilai awal 20% data test.

Hasil Eksperimen

Berdasarkan eksperimen yang dilakukan, berikut adalah hasil yang didapatkan setelah melakukan pengujian terhadap data X_test dan melakukan pengecekan terhadap y_test menggunakan classification_report.

	precision	recall	f1-score	support
True	0.30	0.75	0.43	1336
False	0.95	0.72	0.82	8319
micro avg	0.72	0.72	0.72	9655
macro avg	0.62	0.73	0.62	9655
weighted avg	0.86	0.72	0.76	9655

F1-Score 0.7639928598771132
Accuracy 0.7237700673226307

Seperti yang dapat dilihat pada hasil tersebut bahwa masih terdapat ketimpangan pada hasil yang diperoleh. Hal ini mungkin disebabkan oleh data yang memang sudah *imbalance* sejak semula. Nilai f1-score untuk tebakan True memiliki nilai 0.43 dan f1-score untuk tebakan False memiliki nilai 0.82. Di sisi lain, nilai F1-Score secara menyeluruh adalah 0.76399 dan nilai Accuracy total adalah 0.72377.

Analisis dari Hasil Eksperimen

Berdasarkan hasil eksperimen yang didapatkan, dapat disimpulkan bahwa nilai *classification report* untuk target bernilai True dan False masih terdapat ketimpangan. Hal ini dikarenakan kondisi awal dataset yang sudah *imbalance*. Ketimpangan tersebut sudah sedikit berkurang dikarenakan adanya pelaksanaan *oversampling* sebelum melakukan pelatihan model dan pelaksanaan prediksi.

Nilai F1-Score dan nilai Accuracy sudah menunjukkan nilai yang lumayan tinggi (diatas 0.7). Dengan kata lain, model tersebut dapat sekiranya memprediksi secara akurat sebesar 70%. Hal ini tentunya masih dapat dikembangkan lebih lanjut untuk mendapatkan nilai *metrics* yang lebih baik. Namun, untuk menaikkan nilai *metrics* tersebut tentunya perlu dilakukan perbaikan baik dari sisi pemilihan data, pra-pemrosesan data, maupun pelatihan model.

Kesimpulan

Tujuan dari eksperimen ini adalah untuk memprediksi apakah seseorang memiliki *diabetes* atau tidak berdasarkan data-data yang diberikan. Pelaksanaan prediksi ini dilakukan menggunakan metode pembelajaran mesin dengan pengembangan sebuah model. Pengembangan model ini tentunya melewati langkah dan proses yang panjang, mulai dari analisis data, penanganan analisis data menggunakan teknik tertentu, pelatihan model, dan pelaksanaan prediksi. Model yang kami gunakan dalam mengembangkan program ini adalah model Logistic Regression dan MLP. Hasil prediksi dari kedua model lantas digabungkan dengan menggunakan metode *model stacking* untuk melakukan prediksi terakhir terhadap data *test*. Berdasarkan model yang telah dibuat, model dapat sekiranya secara akurat memprediksi status *diabetes* seseorang sebesar 70% dengan nilai F1-Score secara menyeluruh adalah 0.76399 dan nilai Accuracy total adalah 0.72377.

Pembagian Tugas per Anggota Kelompok

NIM	Nama	Tugas
13521116	Juan Christopher Santoso	Semua (Bagi 2)
13521135	Nicholas Liem	Semua (Bagi 2)