



ANALISIS KLASIFIKASI DIAGNOSIS KARDIOVASKULAR DENGAN FAKTOR GAYA HIDUP DAN DEMOGRAFI: PERBANDINGAN METODE MACHINE LEARNING MENGGUNAKAN REGRESI LOGISTIK, REGRESI LASSO, DAN REGRESI RIDGE

Kelompok 8

Albertus Christian - 00000068921

Evangeline Suciadi - 00000068887

Ivan Bagus Purnomo - 00000069565

Stevani Alexandra Harmareta - 00000069321

LATAR BELAKANG

1

Peran Data Mining dalam Kesehatan:

- Menyediakan alat yang kuat bagi para profesional kesehatan.
- Meksplorasi data besar mengenai kesehatan pasien.

2

Prediksi Risiko Kardiovaskular (CVD):

- Data mining digunakan untuk membuat klasifikasi terkait penyakit kardiovaskular.
- Regresi Logistik, Lasso, dan Ridge membantu mengidentifikasi pasien dengan risiko tinggi dan memungkinkan perancangan intervensi yang tepat.

3

Peran Teknologi Kecerdasan Buatan (AI):

- AI dapat memproses data medis dengan cepat.
- Memberikan prediksi risiko CVD yang lebih akurat.

4

Potensi Pengurangan Dampak Negatif CVD:

- Analisis prediktif dan AI dapat mengurangi dampak negatif penyakit kardiovaskular

RUMUSAN MASALAH

Bagaimana pengaruh variabel gaya hidup, persebaran demografis, dan tes medis terhadap risiko kardiovaskular(CVD)?



Bagaimana model prediktif ini dapat membantu praktisi kesehatan dalam mengidentifikasi pasien dengan risiko tinggi kardiovaskular dan merancang intervensi yang ditargetkan?

Bagaimana perbandingan kinerja regresi logistik, regresi lasso, dan regresi ridge dalam deteksi risiko penyakit kardiovaskular (CVD)?

TUJUAN PENELITIAN

1

Mengembangkan model prediktif yang lebih komprehensif untuk menilai risiko penyakit kardiovaskular (CVD) pada tingkat individu.

2

Meningkatkan aksesibilitas skrining CVD melalui teknologi AI dengan memasukkan informasi tentang pola makan, aktivitas fisik, kebiasaan merokok, paparan polusi udara, dan stres lingkungan lainnya, model ini diharapkan dapat memberikan prediksi yang lebih akurat tentang risiko CVD pada individu

3

Menghasilkan dampak positif dalam pencegahan dan pengelolaan penyakit kardiovaskular di masa depan.

HIPOTESIS

- 1 Faktor gaya hidup, demografi, dan tes medis , memiliki pengaruh signifikan terhadap risiko kardiovaskular.
- 2 Model klasifikasi diagnosis kardiovaskular yang komprehensif dapat meningkatkan kualitas hidup pasien dan mengurangi beban pada sistem kesehatan.
- 3 Regresi Lasso dan Ridge cenderung menghasilkan model diagnosis penyakit kardiovaskular (CVD) yang lebih stabil dan akurat daripada regresi logistik. Ini karena mereka mampu menangani multicollinearity dan overfitting secara lebih efektif, yang dapat meningkatkan konsistensi dan keakuratan prediksi risiko CVD.



TELAAH LITERATUR (RESEARCH GAP)

PENELITIAN	ALGORITMA	POPULASI	LABEL	FEATURES	ACCURACY	JUMLAH SAMPLES
A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL) Mythili T., Dev Mukherji, Nikita Padalia, dan Abhiram Naidu (2021)	Decision Tree, dan Random Forest	920	0 = tidak memiliki penyakit jantung 1 = memiliki penyakit jantung	Umur, Jenis Kelamin, Tekanan Darah, Tingkat Kolesterol, Tingkat Gula Darah, Berat Badan, riwayat penyakit	KNN, DT, RF melalui SVM memperoleh akurasi 100%	736
Heart disease prediction using machine learning algorithms Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain & Preeti Nagrath (2020)	Logistic Regression, dan KNN, Random Forest Classifier	303 Pasien terdiagnosa penyakit jantung	0 = tidak memiliki penyakit jantung 1 = memiliki penyakit jantung	Umur, Jenis Kelamin, Tekanan Darah, Tingkat Kolesterol, Tingkat Gula Darah, Berat Badan, tekanan darah saat istirahat, Hasil tes stres thallium	logistik regresi dan KNN dan mendapatkan akurasi rata-rata 87,5% pada model	212

PENELITIAN	ALGORITMA	POPULASI	LABEL	FEATURES	ACCURACY	JUMLAH SAMPLES
Pengklasifikasian Penyakit Jantung dengan Metode Decision Tree Indrayatna (2020)	KNN, Decision Tree	303 pasien penyakit jantung	0 = tidak memiliki penyakit jantung 1 = punya penyakit jantung	Umur, jenis kelamin, jenis nyeri dada, tekanan darah saat istirahat, jumlah pembuluh darah yang terdeteksi, depresi ST, kemiringan puncak ST	Evaluasi model dengan confusion matrix menghasilkan akurasi 75,4098%	242
Classification models for heart disease prediction using feature selection and PCA Hassani (2020)	CHI-PCA Random Forest, Decision Tree, Logistic Regression, MPC, Naive Bayes, Gradient-Boosted Trees (GBT)	920 pasien dari Cleveland	0 = punya penyakit jantung ; 1,2,3,4 = tingkatan penyakit jantung	ID Pasien, usia, jenis kelamin, nyeri dada, status perokok/tidak, rokok perhari, riwayat penyakit pada keluarga, detak jantung tertinggi, ada atau tidaknya penyakit jantung	CHI-PCA dengan RF memiliki kinerja maksimum 22 kinerja maksimum, dengan akurasi 98,7%	300
HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System Fitriyani (2020)	DBSCAN, SMOTE-ENN, dan XGBOOST	270 pasien penyakit jantung dari dataset I dan 297 pasien dari dataset II	0 = tidak memiliki penyakit jantung 1 = punya penyakit jantung	Umur, Jenis Kelamin, Tekanan Darah, Tingkat Kolesterol, Tingkat Gula Darah, Berat Badan, tekanan darah saat istirahat, Hasil uji stress thallium	Dataset I memiliki akurasi 88% sedangkan dataset II memiliki akurasi 90%	216 dari dataset I dan 237 dari dataset II

- **REGRESI LOGISTIK**

REGRESI LOGISTIK ADALAH METODE STATISTIK YANG DIGUNAKAN UNTUK MEMPREDIKSI VARIABEL DEPENDEN KATEGORI (BIASANYA DENGAN DUA KEMUNGKINAN) BERDASARKAN HUBUNGANNYA DENGAN SATU ATAU LEBIH VARIABEL INDEPENDEN.

- **REGRESI LASSO**

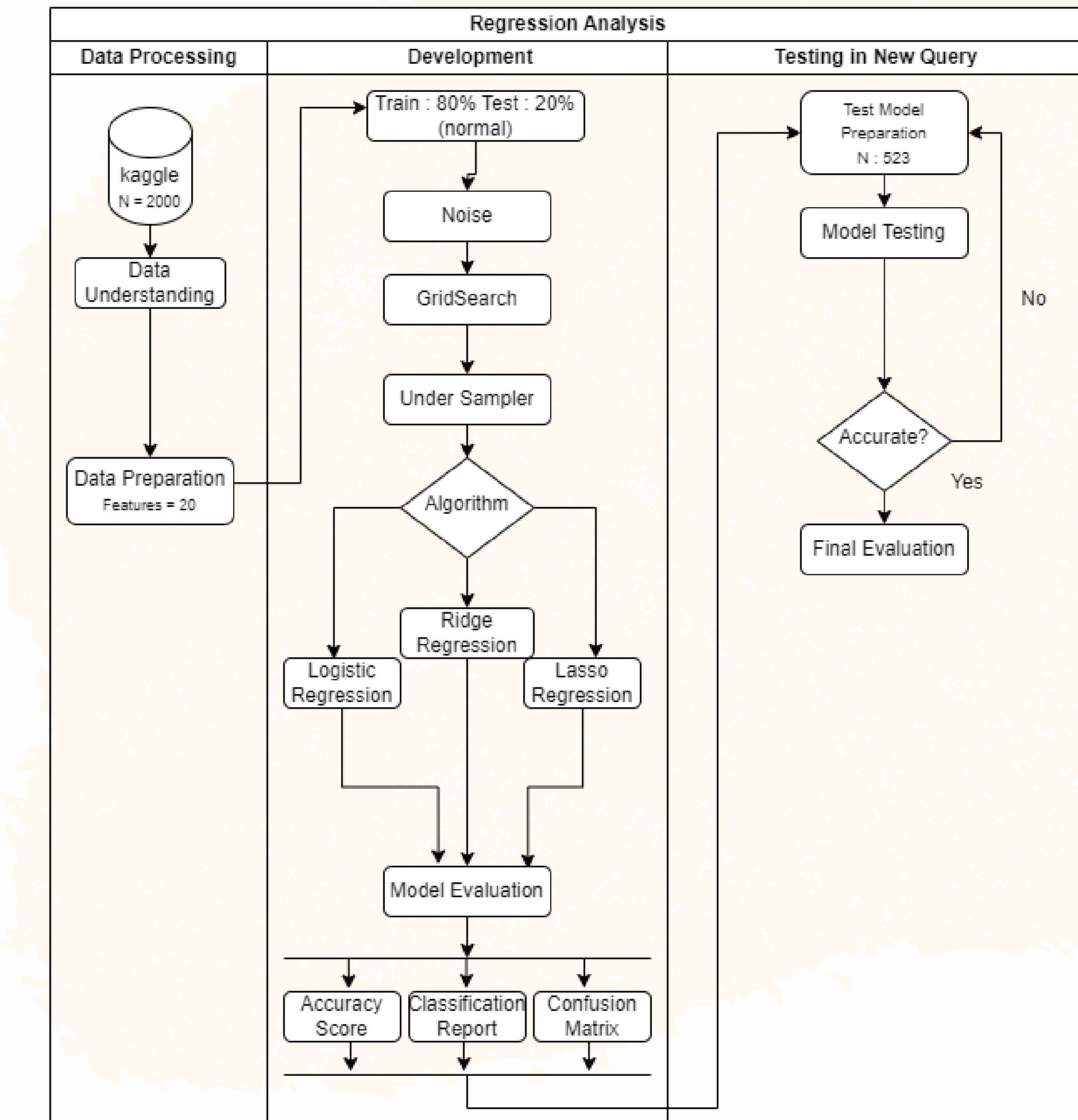
(LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR) ADALAH METODE REGULARISASI YANG DIGUNAKAN UNTUK MENGATASI MASALAH MULTIKOLINEARITAS DAN MENINGKATKAN AKURASI PREDIKSI MODEL REGRESI

- **REGRESI RIDGE,**

JUGA DIKENAL SEBAGAI REGULARISASI TIKHONOV, ADALAH TEKNIK YANG DIGUNAKAN UNTUK MENGANALISIS DATA YANG MENGALAMI MULTIKOLINEARITAS. SEPERTI LASSO, RIDGE MENAMBAHKAN PENALTI KE FUNGSI LOSS, TETAPI DENGAN CARA YANG SEDIKIT BERBEDA

LANDASAN TEORI

METODOLOGI PENELITIAN



1. DATA PROCESSING

Pada penelitian ini, dataset yang digunakan berasal dari survei kesehatan besar yang dijalankan oleh CDC yang disebut BRFSS. Mereka telah mewawancara orang dewasa di AS sejak tahun 1984, dimulai dari 15 negara bagian dan sekarang mencakup seluruh negeri. Survei ini merupakan survei kesehatan terbesar yang sedang berlangsung di dunia, dengan lebih dari 400.000 wawancara setiap tahunnya. Data yang digunakan berasal dari tahun 2020 dan mencakup faktor-faktor yang dapat mempengaruhi penyakit jantung dan memiliki jumlah data sebanyak 31,9795.

<class 'pandas.core.frame.DataFrame'>			
RangeIndex: 1548183 entries, 0 to 1548182			
Data columns (total 28 columns):			
#	Column	Non-Null Count	Dtype
0	HeartDisease	1548183 non-null	int64
1	BMI	1548183 non-null	float64
2	Smoking	1548183 non-null	int64
3	AlcoholDrinking	1548183 non-null	int64
4	Stroke	1548183 non-null	int64
5	PhysicalHealth	1548183 non-null	float64
6	MentalHealth	1548183 non-null	float64
7	DiffWalking	1548183 non-null	int64
8	Sex	1548183 non-null	int64
9	Race	1548183 non-null	int32
10	Diabetic	1548183 non-null	int64
11	PhysicalActivity	1548183 non-null	int64
12	GenHealth	1548183 non-null	int64
13	SleepTime	1548183 non-null	float64
14	Asthma	1548183 non-null	int64
15	KidneyDisease	1548183 non-null	int64
16	SkinCancer	1548183 non-null	int64
17	Age	1548183 non-null	int64
18	cp	1548183 non-null	int64
19	trestbps	1548183 non-null	int64
20	chol	1548183 non-null	int64
21	fbs	1548183 non-null	int64
22	restecg	1548183 non-null	int64
23	thalach	1548183 non-null	int64
24	exang	1548183 non-null	int64
25	oldpeak	1548183 non-null	float64
26	slope	1548183 non-null	int64
27	ca	1548183 non-null	int64
dtypes: float64(5), int32(1), int64(22)			
memory usage: 324.8 MB			

1. DATA PROCESSING

Dataset yang ditampilkan adalah koleksi data medis yang mencakup sebanyak 1,548,183 entri dengan 28 variabel yang berbeda. Setiap kolom dalam dataset ini merepresentasikan atribut medis atau karakteristik dari individu-individu yang terlibat dalam dataset.

Setelah dirapikan, dataset tersebut merupakan kumpulan data besar yang terdiri dari 1,548,183 baris dan 29 kolom dengan column tambahan berupa age group dimana usia pasien dikelompokan menjadi tiga, yaitu kelompok usia 21-39 tahun , kelompok usia 40-60 tahun, dan kelompok usia lebih dari 61 tahun.

Setelah dipilah lebih lanjut lagi, variabel-variabel yang dipilih adalah variabel yang merupakan bagian dari info demografis (demographic_info), dan gaya hidup (lifestyle features).

#	Column	Non-Null Count	Dtype
0	HeartDisease	1548183	non-null
1	BMI	1548183	non-null
2	Smoking	1548183	non-null
3	AlcoholDrinking	1548183	non-null
4	Stroke	1548183	non-null
5	PhysicalHealth	1548183	non-null
6	MentalHealth	1548183	non-null
7	DiffWalking	1548183	non-null
8	Sex	1548183	non-null
9	Race	1548183	non-null
10	Diabetic	1548183	non-null
11	PhysicalActivity	1548183	non-null
12	GenHealth	1548183	non-null
13	SleepTime	1548183	non-null
14	Asthma	1548183	non-null
15	KidneyDisease	1548183	non-null
16	SkinCancer	1548183	non-null
17	Age	1548183	non-null
18	cp	1548183	non-null
19	trestbps	1548183	non-null
20	chol	1548183	non-null
21	fbs	1548183	non-null
22	restecg	1548183	non-null
23	thalach	1548183	non-null
24	exang	1548183	non-null
25	oldpeak	1548183	non-null
26	slope	1548183	non-null
27	ca	1548183	non-null
28	age_group	1548183	non-null

dtypes: category(1), float64(5), int32(1), int64(22)
memory usage: 326.3 MB

```
4 # Lifestyle and Behavior
5 lifestyle_features = ['smoking', 'AlcoholDrinking', 'PhysicalActivity', 'SleepTime']
6
7 # Medical Tests and Indicators
8 medical_tests = ['Stroke', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope', 'ca']
9
10 # Demographic and Personal Information
11 demographic_info = ['Sex', 'Race', 'Age']
12
13 # Functional Abilities
14 functional_abilities = ['DiffWalking']
15
16
17
18 feature_cols = demographic_info + lifestyle_features + medical_tests
```

Clustering Code

1. DATA PROCESSING

Tabel 2. Identifikasi Variabel Faktor Gaya Hidup

Nama Kolom	Definisi
Smoking	indikator apakah seseorang merokok atau tidak (1 untuk "Ya" dan 0 untuk "Tidak")
Alcohol Drinking	indikator mengenai apakah seseorang memiliki kebiasaan konsumsi alkohol (1 untuk "Ya" dan 0 untuk "Tidak")
Physical Activity	Indikator apakah orang tersebut sering melakukan kegiatan olahraga atau tidak (1 untuk "Ya" dan 0 untuk "Tidak")
SleepTime	pengukuran waktu tidur seseorang

Tabel 3. Identifikasi Variabel Faktor Demografis

Nama Kolom	Definisi
Sex	indikator jenis kelamin seseorang (1 untuk "laki-laki" dan 0 untuk "perempuan")
AgeCategory	mengelompokan orang-orang dalam beberapa jangkauan usia
Race	latar belakang ras seseorang

Tabel 5. Identifikasi Variabel Dependen

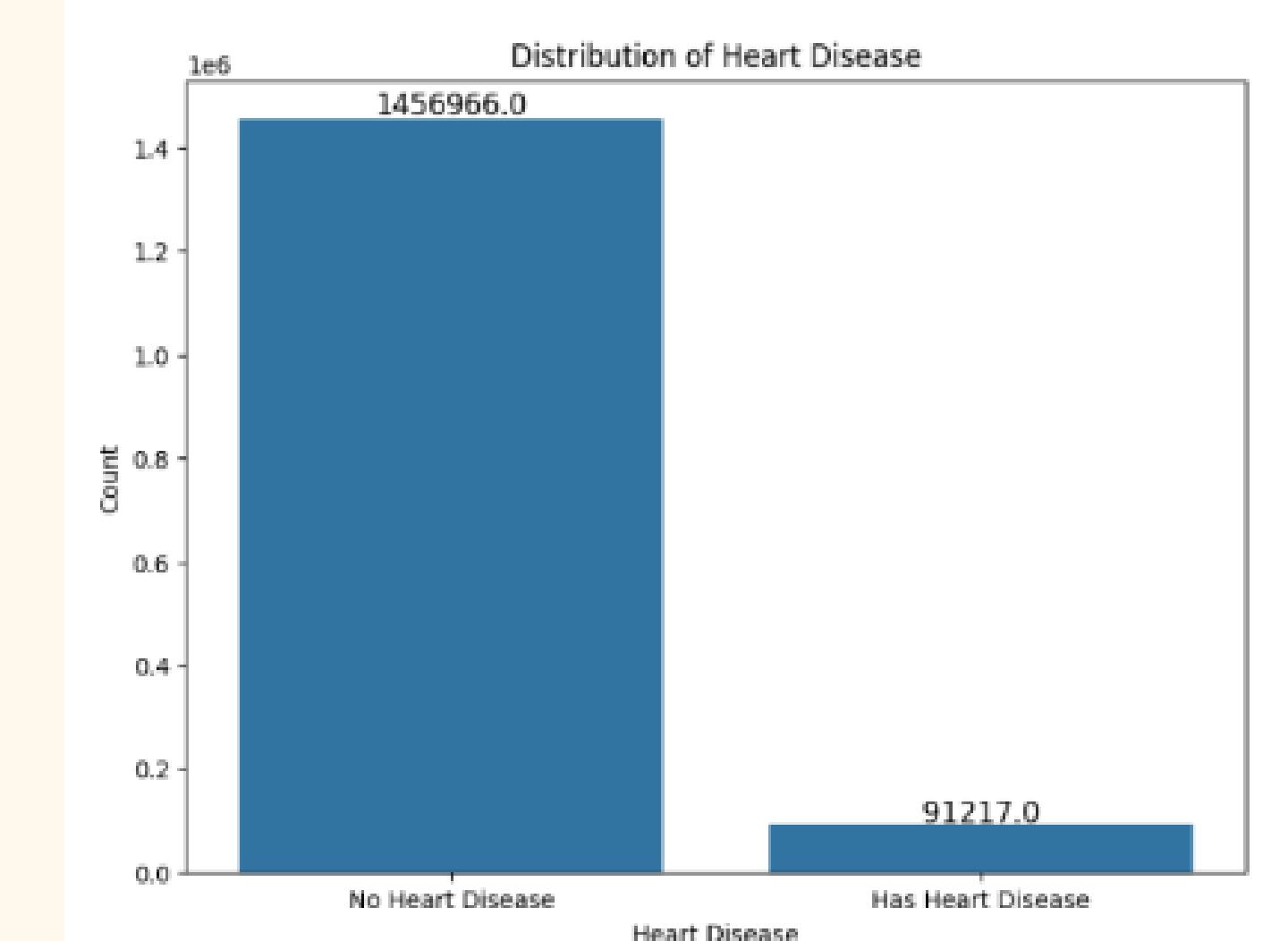
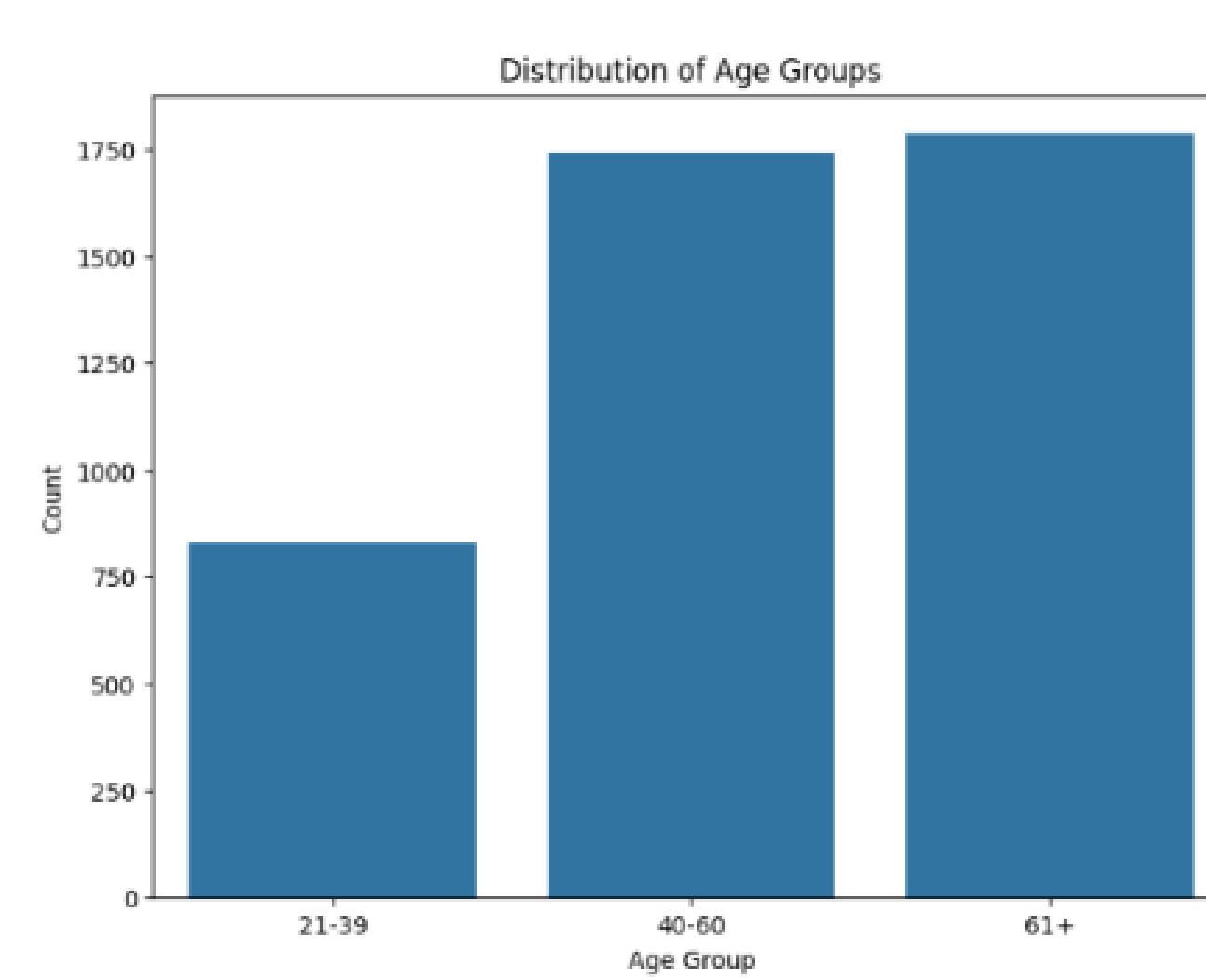
Nama Kolom	Definisi
Heart Disease	menyatakan ada atau tidaknya penyakit kardiovaskular pada pasien (1 untuk "ya" dan 0 untuk "tidak")

Tabel 4. Identifikasi Variabel tes medis

Nama Kolom	Definisi
Heart Disease	Untuk menyatakan hasil akhir, 1 adalah adanya penyakit jantung dan 0 menandakan adanya penyakit jantung
Stroke	Ada atau tidaknya riwayat stroke
cp	4 jenis nyeri dada
trestbps	tekanan darah saat beristirahat
chol	kadar kolesterol dalam mg/dl
fbs	kadar gula darah dalam mg/dl
restecg	hasil dari elektrokardiografi saat beristirahat (dengan nilai 0,1,2)
thalach	detak jantung maksimum yang dicapai
exang	rasa nyeri dada saat berolahraga atau mengalami stres secara emosional
oldpeak	Depresi segmen ST yang diinduksi oleh olahraga dianggap sebagai temuan EKG yang dapat diandalkan untuk diagnosis aterosklerosis koroner obstruktif
slope	kemiringan segmen ST latihan puncak
ca	Jumlah pembuluh darah utama (0-3) yang diwarnai dengan fluoroskopi

Tabel-tabel tersebut menunjukkan 20 variabel saja yang dipilih untuk pemrosesan dan pemodelan data lebih lanjut. Variabel-variabel tersebut dipilih karena berisi faktor-faktor sesuai dengan tujuan penelitian ini, yaitu untuk mengidentifikasi klasifikasi diagnosa kardiovaskular (CDV) melalui tes medis, faktor gaya hidup dan persebaran demografis. Berdasarkan penelitian oleh Widayata [11] , Faktor gaya hidup meliputi kebiasaan merokok, konsumsi alkohol, waktu tidur , kesehatan fisik , kesehatan mental , dan berat badan. Sedangkan untuk faktor demografis terdiri dari jenis kelamin, usia, dan ras. Tes medis berisi tekanan darah , kadar kolesterol, dll.

1. DATA PROCESSING



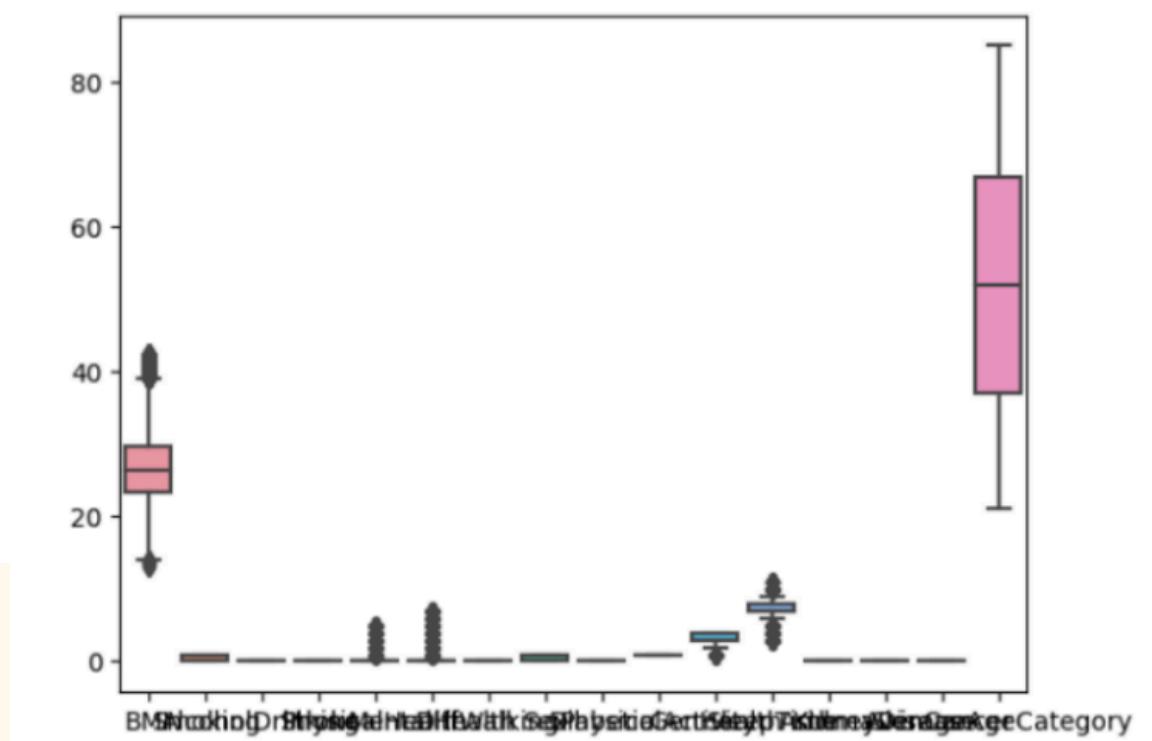
1. DATA PROCESSING

Missing values in each column:

```
HeartDisease      0
BMI              0
Smoking          0
AlcoholDrinking  0
Stroke            0
PhysicalHealth   0
MentalHealth     0
DiffWalking       0
Sex               0
Race              0
Diabetic          0
PhysicalActivity 0
GenHealth          0
SleepTime         0
Asthma            0
KidneyDisease    0
SkinCancer        0
Age               0
```

Hasil pengecekan menunjukkan bahwa tidak ada missing value pada dataset. Selanjutnya adalah tahap pengecekan outlier dengan mengeluarkan variabel non-numerik pada dataset. Pengecekan outlier menggunakan mekanisme nilai interquartile dapat dilihat pada gambar diatas. Dataset tersebut memiliki jumlah pasien pengidap kardiovaskuler sebanyak 3868 orang, sedangkan 89831 pasien tidak memiliki

```
1 # Menghitung IQR dan menghapus outlier dari dataset
2
3 Q1 = df[columns_of_interest].quantile(0.25)
4 Q3 = df[columns_of_interest].quantile(0.75)
5 IQR = Q3 - Q1
6
7 outlier_threshold = 1.5
8
9 outlier_mask = ~((df[columns_of_interest] < (Q1 - outlier_threshold * IQR)) | (df[columns_of_interest] > (Q3 + outlier_threshold * IQR)))
10
11 df_no_outliers = df[outlier_mask]
12
13 sns.boxplot(data=df_no_outliers[columns_of_interest])
14 plt.show()
15
```



```
1 # Pemeriksaan Jumlah Penyakit Jantung
2 heart_disease_counts = df_no_outliers['HeartDisease'].value_counts()
3
4 print("Heart Disease Counts:")
5 print(heart_disease_counts)
```

Heart Disease Counts:
0 89831
1 3868
Name: HeartDisease, dtype: int64

2. DEVELOPMENT

Pada tahapan ini dilakukan proses pembagian data menjadi data train dan data test. Tujuannya adalah untuk mengukur kebaikan model yang terbentuk. Digunakan rasio 80:20 pada penelitian ini, yang artinya 80% data akan digunakan sebagai data train dan 20% sisanya sebagai data test. Pembagian data ini diharapkan mampu melatih model machine learning untuk meningkatkan akurasi.

2. DEVELOPMENT

Logistic Regression

```
In [49]: 1 param_grid_logistic = {  
2     'C': [0.001, 0.01, 0.1, 1, 10, 100],  
3     'penalty': ['l1', 'l2'],  
4     'solver': ['liblinear', 'saga', 'newton-cg', 'lbfgs'],  
5 }
```

```
In [50]: 1 from sklearn.model_selection import GridSearchCV, KFold  
2 modelLogistic = LogisticRegression(max_iter=1000)  
3 kf = KFold(n_splits=5, shuffle=True, random_state=42)
```

```
In [51]: 1 grid_search_logistic = GridSearchCV(modelLogistic, param_grid_logistic,  
2 grid_search_logistic.fit(X_train, y_train)
```

```
[2]: GridSearchCV  
  - estimator: LogisticRegression  
    - LogisticRegression
```

```
In [52]: 1 best_model = grid_search_logistic.best_estimator_
```

```
In [53]: 1 y_pred = best_model.predict(X_test)
```

Lasso Regression

```
In [49]: 1 # Lasso Regression  
2 from sklearn.linear_model import Lasso
```

```
In [50]: 1 param_grid_lasso = {  
2     'alpha': [0.001, 0.01, 0.1, 1, 10, 100],  
3     'max_iter': [1000, 2000, 3000],  
4     'tol': [0.0001, 0.001, 0.01],  
5     'selection': ['cyclic', 'random']  
6 }
```

```
In [51]: 1 # Inisialisasi model lasso dengan alpha (parameter regularisasi)  
2 lasso_model = Lasso(alpha=1.0)  
3 lasso_model.fit(X_train, y_train)
```

```
Out[51]: Lasso  
Lasso()
```

```
In [52]: 1 kf = KFold(n_splits=5, shuffle=True, random_state=42)
```

```
In [53]: 1 grid_search_lasso = GridSearchCV(lasso_model, param_grid_lasso, cv=kf, s
```

```
Out[53]: GridSearchCV  
  - estimator: Lasso  
    - Lasso
```

```
In [54]: 1 best_model_lasso = grid_search_lasso.best_estimator_
```

```
In [55]: 1 # Prediksi nilai target pada data uji  
2 y_pred_lasso = best_model_lasso.predict(X_test)  
3  
4 # Mengubah prediksi menjadi kelas biner berdasarkan threshold 0.5  
5 y_pred_lasso_binary = (y_pred_lasso > 0.5).astype(int)
```

2. DEVELOPMENT

```
Ridge Regression

In [60]: M 1 param_grid_ridge = {
2     'alpha': [0.001, 0.01, 0.1, 1, 10, 100],
3     'max_iter': [1000, 2000, 3000],
4     'tol': [0.0001, 0.001, 0.01],
5     'solver': ['auto', 'svd', 'cholesky', 'lsqr', 'sparse_cg', 'sag', 's
6 }

In [61]: M 1 from sklearn.linear_model import Ridge
2
3 # Buat dan sesuaikan model Ridge dengan alpha (parameter regularisasi)
4 model_ridge = Ridge(alpha=1.0)
5 model_ridge.fit(X_train, y_train)

Out[61]: > Ridge
Ridge()

In [62]: M 1 kf = KFold(n_splits=5, shuffle=True, random_state=42)

In [63]: M 1 grid_search_ridge = GridSearchCV(model_ridge, param_grid_ridge, cv=kf, s
2 grid_search_ridge.fit(X_train, y_train)

Out[63]: > GridSearchCV
> estimator: Ridge
    > Ridge

In [64]: M 1 best_model_ridge = grid_search_ridge.best_estimator_

In [65]: M 1 # Prediksi nilai target pada data uji menggunakan model Ridge
2 y_pred_ridge = best_model_ridge.predict(X_test)

In [66]: M 1 # Mengubah prediksi menjadi kelas biner berdasarkan threshold 0.5
2 y_pred_Ridge_Binary = (y_pred_ridge > 0.5).astype(int)
```

Proses pencarian grid yang dimplementasikan pada analisa regresi ini mengeksplorasi berbagai kombinasi hiperparameter ini dan mengevaluasi setiap kombinasi menggunakan validasi silang 5 kali lipat. Teknik ini membagi data pelatihan menjadi lima fold, melatih model pada empat fold, dan mengujinya pada fold yang tersisa. Proses ini diulangi sebanyak lima kali, untuk memastikan kinerja model yang kuat di berbagai subset data. Pada akhirnya, pencarian grid memilih kombinasi hyperparameter yang menghasilkan performa terbaik sesuai dengan matrik penilaian yang dipilih (ROC AUC dalam kasus ini). Pendekatan ini membantu untuk menemukan model regresi logistik yang cenderung menggeneralisasi dengan baik pada data yang tidak terlihat, membuat prediksi yang akurat pada contoh-contoh yang akan datang.

2. DEVELOPMENT

Tabel Perbandingan Sampel Akurasi

Model	Ukuran Training (rasio)	Ukuran Tes (rasio)	Ketepatan
Regresi logistik	80%	20%	0,9825
	70%	30%	0,9814
	60%	40%	0,9857
Regresi LASSO	80%	20%	0,9890
	70%	30%	0,8575
	60%	40%	0,8975
Regresi Punggung Bukit	80%	20%	0,9625.
	70%	30%	0,96667
	60%	40%	0,99812

Tabel ini memberikan analisis komparatif akurasi pengujian tiga model regresi yang berbeda. Untuk Regresi Logistik, keakuratannya tetap tinggi secara konsisten di berbagai bagian pengujian pelatihan: 0,9825 untuk 80%-20%, 0,9814 untuk 70%-30%, dan 0,9857 untuk 60%-40%. Model ini berperforma baik terlepas dari pemisahan datanya, dengan akurasi yang sedikit meningkat seiring bertambahnya ukuran pengujian.

Sebaliknya, Regresi Lasso menunjukkan akurasi tinggi dengan data pelatihan 80% sebesar 0,9890, namun terjadi penurunan performa yang signifikan ketika ukuran pelatihan dikurangi, mencapai 0,8575 untuk 70%-30% dan 0,8975 untuk 60%-40%.

Regresi Ridge, di sisi lain, menunjukkan akurasi yang tinggi dan meningkat seiring bertambahnya ukuran pengujian: 0,9625 untuk 80%-20%, 0,96667 untuk 70%-30%, dan 0,99812 yang mengesankan untuk 60%-40%.

Model ini berperforma baik secara konsisten dan menggeneralisasi secara efektif dengan kumpulan data pengujian yang lebih besar. Perbandingan ini menggarisbawahi pentingnya memilih model yang tepat dan pemisahan pengujian pelatihan, karena model yang berbeda memberikan respons yang berbeda terhadap perubahan jumlah data pelatihan dan pengujian.

3. NEW QUERY TESTING

Tabel tersebut membandingkan akurasi pelatihan dan pengujian Regresi Logistik, Regresi Lasso, dan Regresi Ridge, yang menyoroti peningkatan masing-masing dari pelatihan hingga pengujian.

Model	Akurasi Pelatihan	Akurasi Pengujian	Peningkatan
Regresi logistik	1	0,91169451	0,08830549
Regresi Lasso	0,969375	0,99761336	-0,02823836
Regresi Punggung Bukit	0,976875	0,99761336	-0,02073836

Regresi Logistik mencapai akurasi pelatihan yang sempurna tetapi turun menjadi sekitar 91,17% dalam akurasi pengujian, sehingga menghasilkan peningkatan sebesar 0,0883, yang menunjukkan beberapa tingkat overfitting.

Sebaliknya, Regresi Lasso memiliki akurasi pelatihan sekitar 96,94% dan akurasi pengujian lebih tinggi yaitu 99,76%. Peningkatan negatif sebesar -0,0282 menunjukkan bahwa Regresi Lasso berkinerja lebih baik pada data pengujian, menunjukkan generalisasi yang kuat karena regularisasi.

Demikian pula, Regresi Ridge menunjukkan akurasi pelatihan sekitar 97,69% dan akurasi pengujian 99,76%, dengan peningkatan -0,0207, mencerminkan generalisasi yang sangat baik.

Di antara model-model tersebut, Regresi Lasso dan Ridge lebih unggul daripada Regresi Logistik, dengan Regresi Ridge sedikit lebih baik daripada Regresi Lasso karena peningkatan negatifnya yang lebih kecil, menunjukkan bahwa model tersebut menggeneralisasi sedikit lebih baik terhadap data yang tidak terlihat.

KESIMPULAN

Sebagai kesimpulan, analisis komparatif model Regresi Logistik, Lasso, dan Ridge Regression untuk skrining penyakit jantung menggarisbawahi keunggulan Regresi Ridge karena kemampuan generalisasinya yang luar biasa dan kinerjanya yang konsisten. Regresi Logistik, meskipun memiliki akurasi pelatihan yang sempurna, menunjukkan tanda-tanda overfitting dengan penurunan yang mencolok dalam akurasi pengujian. Di sisi lain, baik Lasso dan Ridge Regression menunjukkan peningkatan negatif, yang mengindikasikan generalisasi yang kuat yang difasilitasi oleh teknik regularisasi, dengan Ridge Regression sedikit mengungguli Lasso.

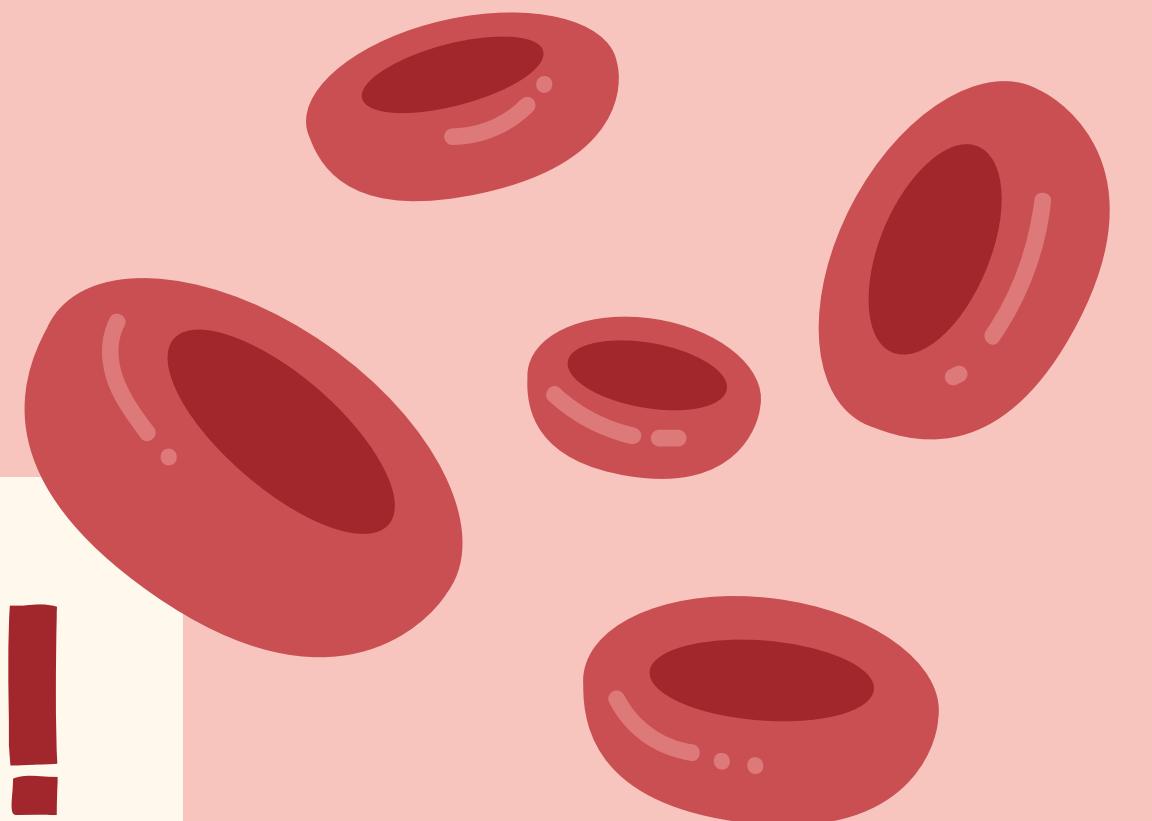
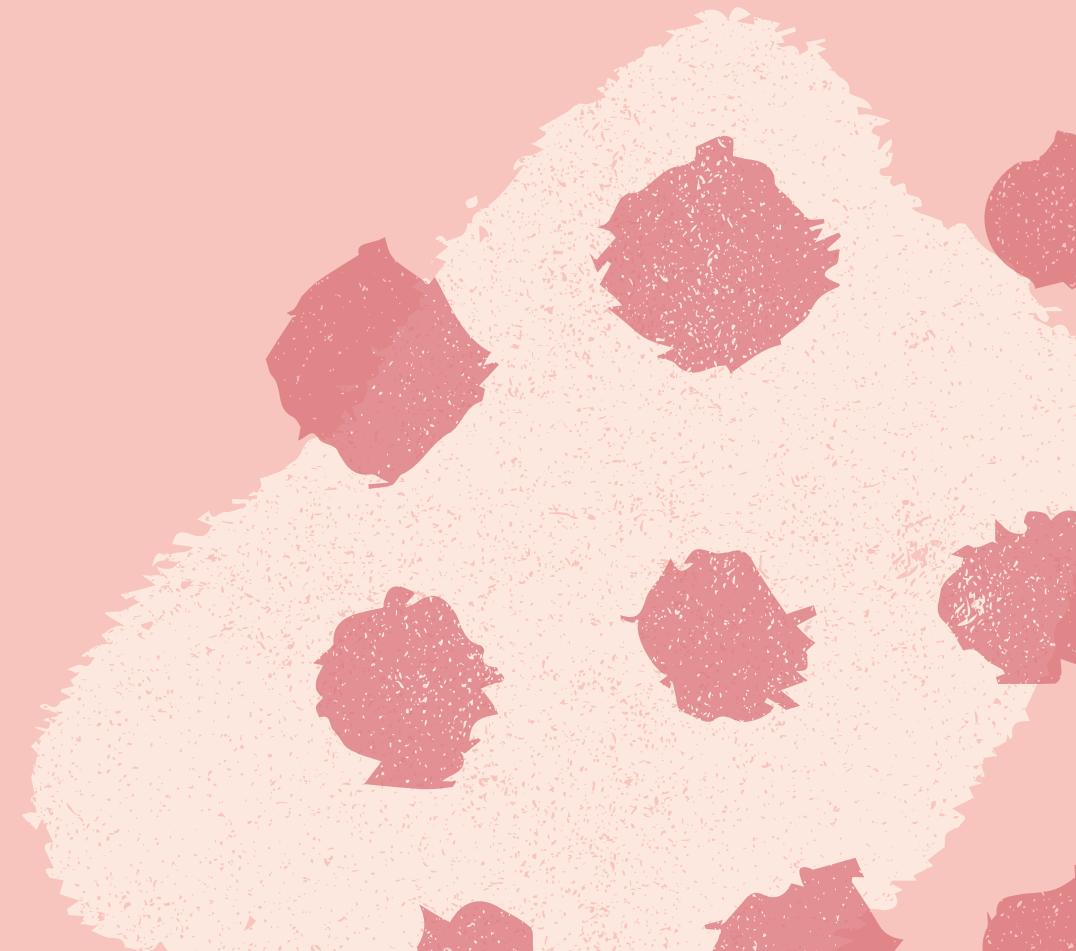
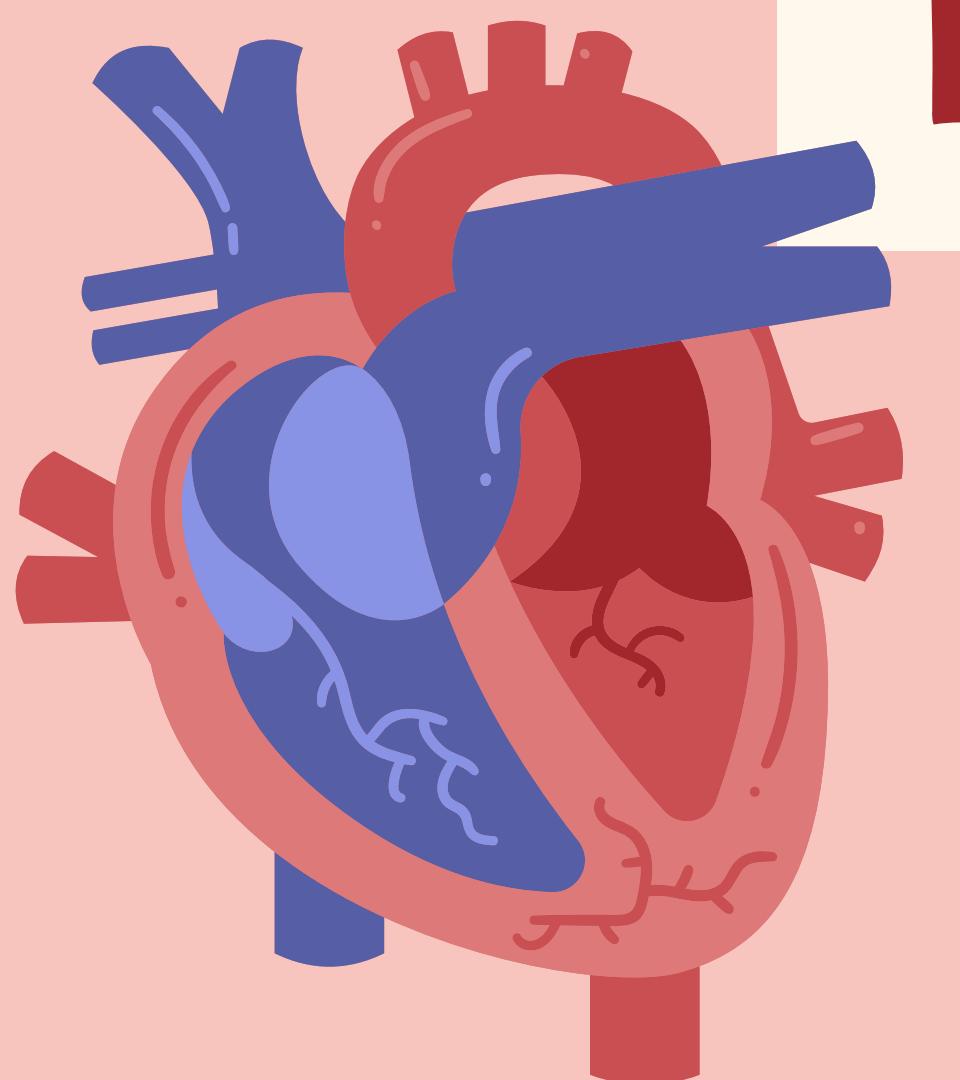
Untuk lebih meningkatkan akurasi prediksi dari model-model ini, beberapa strategi harus diterapkan. Teknik pembelajaran ensemble, yang menggabungkan kekuatan beberapa model regresi, dapat menghasilkan prediksi yang lebih akurat dan kuat. Rekayasa fitur sangat penting untuk memilih dan memproses fitur data yang paling relevan, sehingga meningkatkan efektivitas model. Penyetelan hiperparameter yang optimal dapat menyempurnakan kinerja model, memastikan akurasi yang lebih baik. Selain itu, memperluas set data yang digunakan untuk pelatihan dengan data pasien yang lebih komprehensif dapat secara signifikan memperkuat prediksi model.



KESIMPULAN

Arah penelitian di masa depan harus fokus pada pengumpulan dan analisis data pasien yang luas untuk menyempurnakan dan memvalidasi model prediktif untuk risiko penyakit kardiovaskular (CVD). Selain itu, mengeksplorasi teknik AI yang canggih seperti deep learning dapat meningkatkan akurasi prediksi lebih lanjut, menawarkan pendekatan yang lebih canggih untuk skrining penyakit jantung. Melalui upaya gabungan ini, pengembangan model yang andal dan sangat akurat untuk penilaian risiko CVD dapat diwujudkan, yang pada akhirnya berkontribusi pada hasil yang lebih baik bagi pasien dan tindakan perawatan kesehatan preventif.





THANK YOU !