PENGEMBANGAN MODEL ANALISIS PREDIKTIF DARI BIG DATA UNTUK PREDIKSI PENYAKIT JANTUNG MENGGUNAKAN TEKNIK MACHINE LEARNING

PROJEK UJIAN AKHIR SEMESTER BIG DATA

Anggota Kelompok

Irza Ramira Putra	1810511100
Deo Haganta Depari	1810511104
Nadhifa Zhafira	1810511111
Quina Alifa	1810511115

Pendahuluan

Center for Disease Control and Prevention menulis bahwa **penyakit jantung merupakan penyebab utama** meninggalnya pria, wanita dan orang-orang dari sebagian besar kelompok ras dan etnis di Amerika Serikat.

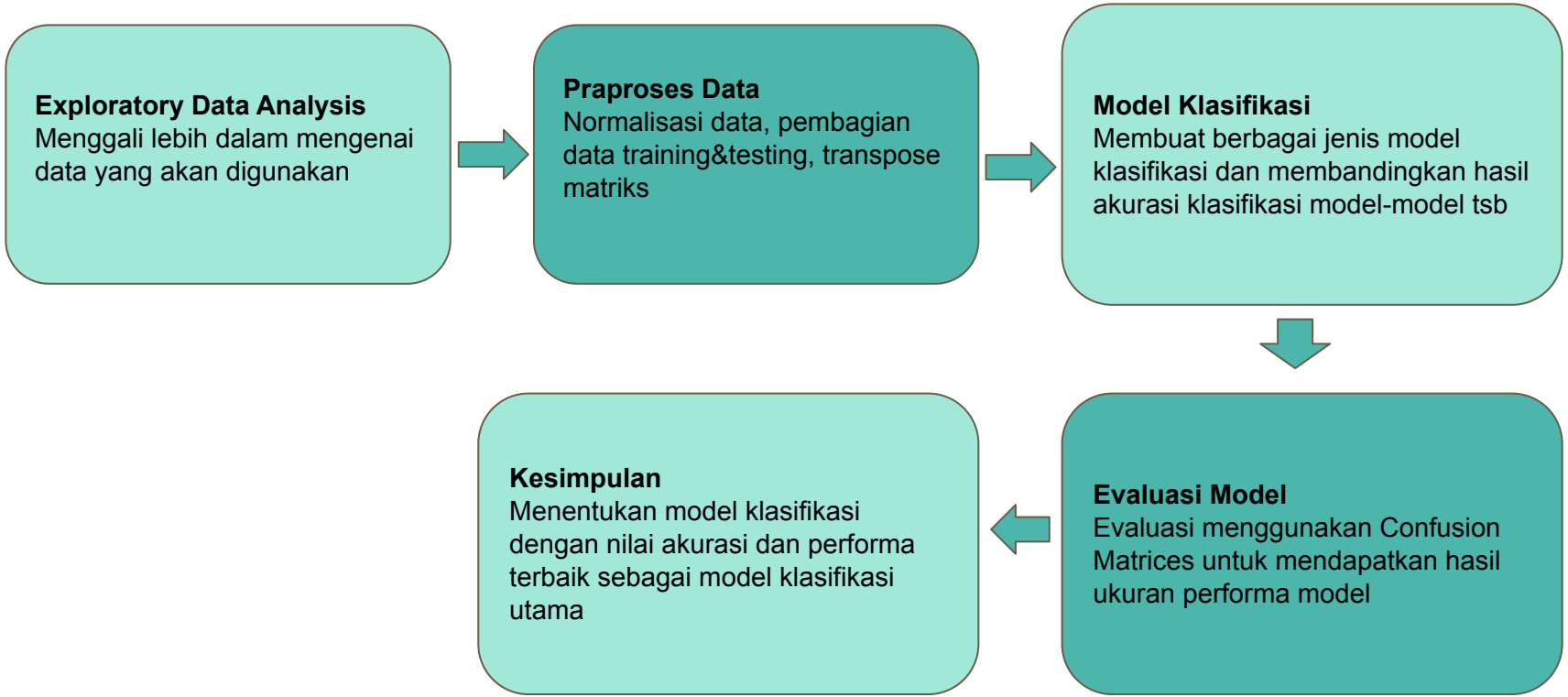
Berdasarkan data Riset Kesehatan Dasar (Riskesdas) tahun 2018, angka kejadian penyakit jantung dan pembuluh darah semakin meningkat dari tahun ke tahun. Setidaknya, 15 dari 1000 orang, atau sekitar 4,2 juta individu di Indonesia menderita penyakit jantung.

Diagnosis

Dengan adanya digitalisasi informasi medis, data medis semakin banyak dan perlu dikonversi menjadi informasi yang berguna dan dapat ditindaklanjuti.

Penelitian ini dilakukan dengan mengolah data berdasarkan atributnya untuk mendiagnosis pasien penyakit jantung. Hasil akhir dari penelitian ini adalah perbandingan nilai akurasi dan performa dari beberapa model klasifikasi, sehingga dapat menentukan model yang tepat

Metode Penelitian

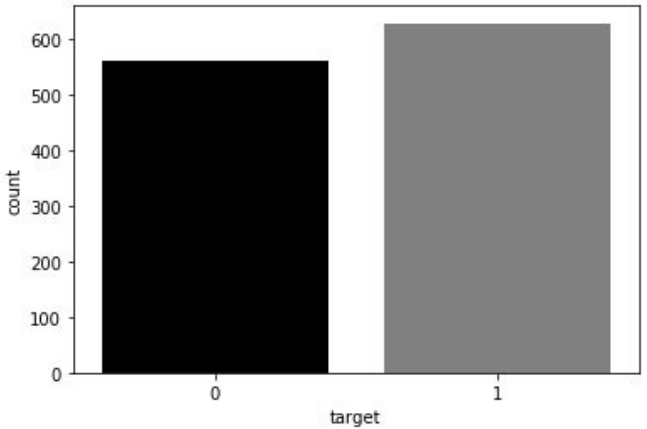


Pembahasan

Exploratory Data Analysis

Penelitian ini menggunakan dataset ***HEART DISEASE DATASET (COMPREHENSIVE)*** yang telah dimodifikasi dimana data terdiri dari 12 atribut. Data tersebut didapat dari ***IEEE***.

#	Column	Count	Non-Null	Dtype
0	age	1190	non-null	int64
1	sex	1190	non-null	int64
2	cp	1190	non-null	int64
3	trestbps	1190	non-null	int64
4	chol	1190	non-null	int64
5	fbs	1190	non-null	int64
6	restecg	1190	non-null	int64
7	thalach	1190	non-null	int64
8	exang	1190	non-null	int64
9	oldpeak	1190	non-null	float64
10	slope	1190	non-null	int64
11	target	1190	non-null	int64



Pasien yang memiliki Penyakit Jantung **= 52.82%**

Pasien yang tidak memiliki Penyakit Jantung **= 47.18%**

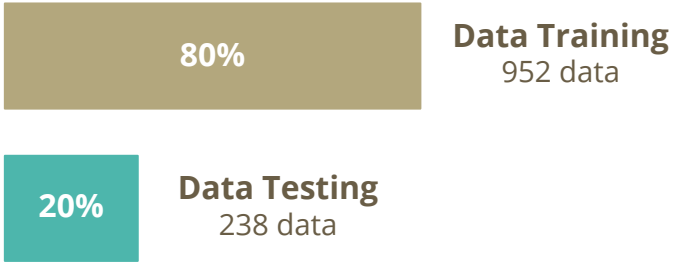
Praproses Data

- **Mengganti atribut** *slope* pada data frame dengan *dummy variable*
- **Membagi data** menjadi data fitur dan data kelas
- Melakukan **Normalisasi Data** dengan rumus

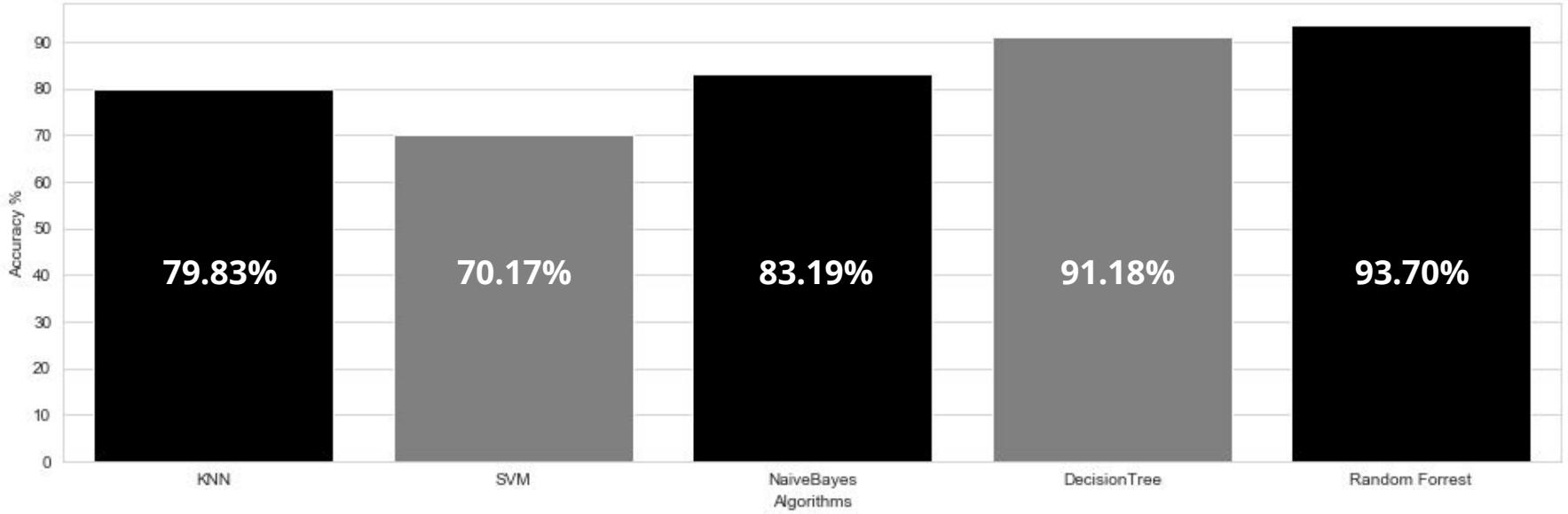
$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Pembagian Data

Dilakukan pemisahan data untuk training dan testing.



Pembuatan Model



Evaluasi Hasil

Proses selanjutnya yaitu evaluasi dari model-model yang sudah dihasilkan. Evaluasi yang dihasilkan adalah ***Precision, Recall,*** dan ***F1-Score*** dengan ***Confusion Matrix.***

Tabel 1. Hasil Evaluasi model KNN dengan Data Training

	0	1	Accuracy	Macro Avg	Weighted Avg
Precision	1.0	1.0	1.0	1.0	1.0
Recall	1.0	1.0	1.0	1.0	1.0
F1-Score	1.0	1.0	1.0	1.0	1.0
Support	459.0	492.0	1.0	951.0	951.0

Tabel 2. Hasil Evaluasi model KNN dengan Data Testing

	0	1	Accuracy	Macro Avg	Weighted Avg
Precision	0.76	0.82	0.80	0.79	0.80
Recall	0.76	0.82	0.80	0.79	0.80
F1-Score	0.76	0.82	0.80	0.79	0.80
Support	102.00	136.00	0.80	238.00	238.00

Tabel 3. Confusion Matrix dengan Model KNN

Train Data

	Actually Positive	Actually Negative
Predicted Positive	459	0
Predicted Negative	0	492

Test Data

	Actually Positive	Actually Negative
Predicted Positive	78	24
Predicted Negative	24	112

Tabel 4. Hasil Evaluasi model SVM dengan Data Training

	0	1	Accuracy	Macro Avg	Weighted Avg
Precision	0.68	0.75	0.71	0.71	0.72
Recall	0.76	0.67	0.71	0.71	0.72
F1-Score	0.72	0.71	0.71	0.71	0.72
Support	459.00	492.00	0.71	951.00	951.00

Tabel 5. Hasil Evaluasi model SVM dengan Data Testing

	0	1	Accuracy	Macro Avg	Weighted Avg
Precision	0.62	0.79	0.70	0.70	0.71
Recall	0.76	0.65	0.70	0.71	0.70
F1-Score	0.68	0.71	0.70	0.70	0.70
Support	102.00	136.00	0.70	238.00	238.00

Tabel 6. Confusion Matrix model SVM

Train Data

	Actually Positive	Actually Negative
Predicted Positive	350	109
Predicted Negative	160	332

Test Data

	Actually Positive	Actually Negative
Predicted Positive	78	24
Predicted Negative	48	88

Evaluasi Hasil

Tabel 7. Hasil Evaluasi model Naive Bayes dengan Data Training

	0	1	Accuracy	Macro Avg	Weighted Avg
Precision	0.82	0.86	0.84	0.84	0.84
Recall	0.85	0.82	0.84	0.84	0.84
F1-Score	0.83	0.84	0.84	0.84	0.84
Support	459.00	492.00	0.84	951.00	951.00

Tabel 8. Hasil Evaluasi model Naive Bayes dengan Data Testing

	0	1	Accuracy	Macro Avg	Weighted Avg
Precision	0.83	0.90	0.87	0.87	0.87
Recall	0.87	0.87	0.87	0.87	0.87
F1-Score	0.85	0.88	0.87	0.87	0.87
Support	102.00	136.00	0.87	238.00	238.00

Tabel 9. Confusion Matrix model Naive Bayes

Train Data

	Actually Positive	Actually Negative
Predicted Positive	391	68
Predicted Negative	88	404

Test Data

	Actually Positive	Actually Negative
Predicted Positive	89	13
Predicted Negative	18	118

Tabel 10. Hasil Evaluasi model Decision Tree dengan Data Training

	0	1	Accuracy	Macro Avg	Weighted Avg
Precision	1.0	1.0	1.0	1.0	1.0
Recall	1.0	1.0	1.0	1.0	1.0
F1-Score	1.0	1.0	1.0	1.0	1.0
Support	459.0	492.0	1.0	951.0	951.0

Tabel 11. Hasil Evaluasi model Decision Tree dengan Data Testing

	0	1	Accuracy	Macro Avg	Weighted Avg
Precision	0.89	0.90	0.89	0.89	0.89
Recall	0.86	0.92	0.89	0.89	0.89
F1-Score	0.88	0.91	0.89	0.89	0.89
Support	102.00	136.00	0.89	238.00	238.00

Tabel 12. Confusion Matrix model Decision Tree

Train Data

	Actually Positive	Actually Negative
Predicted Positive	459	0
Predicted Negative	0	492

Test Data

	Actually Positive	Actually Negative
Predicted Positive	88	14
Predicted Negative	11	125

Evaluasi Hasil

Tabel 13. Hasil Evaluasi model Random Forest dengan Data Training

	0	1	Accuracy	Macro Avg	Weighted Avg
Precision	1.0	1.0	1.0	1.0	1.0
Recall	1.0	1.0	1.0	1.0	1.0
F1-Score	1.0	1.0	1.0	1.0	1.0
Support	459.0	492.0	1.0	951.0	951.0

Tabel 14. Hasil Evaluasi model Random Forest dengan Data Testing

	0	1	Accuracy	Macro Avg	Weighted Avg
Precision	0.97	0.92	0.94	0.95	0.94
Recall	0.90	0.98	0.94	0.94	0.94
F1-Score	0.93	0.95	0.94	0.94	0.94
Support	102.00	136.00	0.94	238.00	238.00

Tabel 15. Confusion Matrix model Random Forest

Train Data

	Actually Positive	Actually Negative
Predicted Positive	459	0
Predicted Negative	0	492

Test Data

	Actually Positive	Actually Negative
Predicted Positive	91	11
Predicted Negative	3	133

Penutup

Kesimpulan

Berdasarkan hasil penelitian, **model Random Forest yang paling tepat untuk digunakan** untuk diagnosis pasien penyakit jantung. Dari hasil evaluasi model, nilai akurasi dari tinggi ke rendah yaitu:

Saran

- Dapat dibuat dimana **hasil prediksi dari data pasien dapat disimpan** ke dalam dataset utama yang tentunya sudah di validasi oleh ahli medis dan juga sudah diberikan izin oleh pasien untuk digunakan, sehingga data akan menjadi lebih besar dan bervariasi.
- Diharapkan **dapat digunakan secara langsung** oleh para ahli medis dan dapat membantu ahli medis.

Terima Kasih