ANALISIS PENGARUH *FEATURE SELECTION* PADA MODEL *HYBRID MACHINE LEARNING* UNTUK PREDIKSI PENYAKIT JANTUNG

SKRIPSI

Program Studi Sarjana Informatika
Jurusan Informatika

Oleh:

FIKRI AL JAUZI

NIM D1041221075



FAKULTAS TEKNIK

UNIVERSITAS TANJUNGPURA

PONTIANAK

2026

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini:

Nama : Fikri Al Jauzi

NIM : D1041221075

menyatakan bahwa dalam SKRIPSI yang berjudul "Analisis Pengaruh *Feature selection* Pada Model *Hybrid Machine Learning* Untuk Prediksi Penyakit Jantung" tidak terdapat karya yang pernah diajukan untuk memperoleh gelar sarjana di suatu perguruan tinggi manapun. Sepanjang pengetahuan Saya, tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam Daftar Rujukan.

Demikian pernyataan ini dibuat dengan sebenar-benarnya. Saya sanggup menerima konsekuensi akademis dan hukum di kemudian hari apabila pernyataan yang dibuat ini tidak benar.

Pontianak, 6 Juli 2026

Fikri Al Jauzi

NIM D1041221075

KEMENTERIAN PENDIDIKAN TINGGI, SAINS, DAN TEKNOLOGI
UNIVERSITAS TANJUNGPURA
FAKULTAS TEKNIK
Jalan Prof. Dr. H. Hadari Nawawi Pontianak 78124
Telepon (0561) 740186, WA: +6282152280907
Email : ft@untan.ac.id Website : http://teknik.untan.ac.id

HALAMAN PENGESAHAN

ANALISIS PENGARUH *FEATURE SELECTION* PADA MODEL *HYBRID MACHINE LEARNING* UNTUK PREDIKSI PENYAKIT JANTUNG

Jurusan Informatika
Program Studi Sarjana Informatika

Oleh:
Fikri Al Jauzi
NIM. D1041221075

Telah dipertahankan di depan Penguji Skripsi pada tanggal 6 Juli 2026
dan diterima sebagai salah satu persyaratan untuk memperoleh gelar sarjana

Susunan Penguji Skripsi

Dosen Pembimbing Utama	: Dr. Arif Bijaksana Putra Negara, S.T., M.T. NIP. 197208081998021002
Dosen Pembimbing Kedua	: Niken Candraningrum, S.T., M.Cs. NIP. 199309272022032010
Dosen Penguji Utama	: Tursina, S.T., M.Cs. NIP. 197801152002122003
Dosen Penguji Kedua	: Khairul Hafidh, S.T., M.Kom. NIP. 199202232022031003

Pontianak, 6 Juli 2026

Dekan

Dr. Ing. Ir. Slamet Widodo, MT, IPM
NIP. 196412231992031002

Pembimbing Utama

Dr. Arif Bijaksana Putra Negara, S.T., M.T.
NIP. 197208081998021002

HALAMAN PERSEMBAHAN

Dengan penuh rasa Syukur kepada Allah Swt atas segala rahmat, kasih, dan penyertaan-Nya, karya sederhana ini penulis persembahkan kepada :

1. Kedua orang tua, yang selalu menjadi sumber kekuatan terbesar dalam kehidupan penulis. Terima kasih atas segala doa yang tidak pernah putus, pengorbanan yang tak terhitung, serta dukungan moral dan material yang diberikan dengan penuh kasih sayang. Setiap pencapaian yang diraih penulis tidak terlepas dari peran dan perjuangan kalian.
2. Keluarga besar penulis, yang senantiasa memberikan perhatian, motivasi, serta doa sehingga penulis mampu melewati berbagai tantangan selama menjalani masa perkuliahan.
3. Para dosen Program Studi Informatika Fakultas Teknik Universitas Tanjungpura, khususnya dosen pembimbing dan dosen penguji yang telah meluangkan waktu, tenaga, dan pemikirannya untuk memberikan arahan, masukan, serta ilmu yang sangat berharga selama proses penyusunan skripsi ini.
4. Rekan-rekan seperjuangan Informatika Angkatan 2022, yang telah menjadi teman belajar, berdiskusi, berbagi pengalaman, dan saling mendukung dalam menyelesaikan berbagai tantangan akademik selama masa perkuliahan.
5. Keluarga besar Himpunan Mahasiswa Informatika (HMIF) Fakultas Teknik UNTAN dan Badan Eksekutif Mahasiswa Fakultas Teknik UNTAN, yang telah memberikan ruang bagi penulis untuk berkembang, belajar mengenai kepemimpinan, kerja sama tim, serta pengabdian kepada organisasi dan mahasiswa.
6. Sahabat, teman dekat, dan seluruh rekan seperjalanan, yang selalu hadir dalam berbagai kondisi, memberikan semangat, dukungan, serta menjadi tempat bertukar pikiran selama proses penyelesaian studi.

Semua pihak yang telah membantu, baik secara langsung maupun tidak langsung, yang tidak dapat disebutkan satu per satu. Setiap bantuan, doa, dan kebaikan yang diberikan memiliki arti yang sangat besar dalam perjalanan penulis hingga menyelesaikan skripsi ini.

PRAKATA

Segala puji bagi Allah SWT atas rahmat-Nya sehingga penulis dapat menyelesaikan penelitian serta penyusunan skripsi yang berjudul "Analisis Pengaruh *Feature selection* Pada Model *Hybrid Machine Learning* Untuk Prediksi Penyakit Jantung" dengan baik dan tepat waktu.

Skripsi ini membahas penerapan metode *Machine Learning* dalam memprediksi penyakit jantung. Penelitian ini juga mengkaji pengaruh implementasi *feature selection* menggunakan pengujian *friedman* sehingga diperoleh pemahaman yang lebih mendalam perbandingan skenario – skenario yang akan dibandingkan.

Secara khusus penulis menyampaikan ucapan terimakasih yang sebesar-besarya kepada Bapak Dr. Arif Bijaksana Putra Negara, S.T., M.T. dan Ibu Niken Candraningrum, S.T., M.Cs. selaku dosen Pembimbing Utama dan dosen Pembimbing Kedua, serta Ibu Tursina, S.T., M.Cs. dan Bapak Khairul Hafidh, S.T., M.Kom. selaku dosen Penguji Utama dan dosen Penguji Kedua yang telah memberikan arahan, bimbingan, serta masukan selama proses penelitian dan penyusunan skripsi ini. Ucapan terimakasih juga disampaikan kepada seluruh pihak yang telah memberikan dukungan baik secara langsung maupun tidak langsung sehingga penelitian ini dapat terselesaikan dengan baik.

Penulis menyadari bahwa penelitian ini masih memiliki keterbatasan baik dari segi metode maupun data yang digunakan. Oleh karena itu, penulis sangat mengharapkan kritik dan saran yang membangun untuk penyempurnaan penelitian di masa yang akan datang. Penulis berharap semoga penelitian ini dapat memberikan manfaat serta kontribusi dalam pengembangan ilmu pengetahuan khususnya dalam bidang analisis data dan prediksi harga saham.

Pontianak, 6 Juli 2022

Penulis,

ABSTRAK

Penyakit jantung merupakan salah satu penyebab utama kematian di dunia sehingga diperlukan sistem prediksi yang akurat dan efisien untuk mendukung deteksi dini. Salah satu pendekatan yang banyak digunakan adalah *Machine Learning*, termasuk model *Hybrid* yang menggabungkan *Extra Trees Classifier (ETC)* dan *Extreme Gradient Boosting (XGBoost)*. Penelitian ini bertujuan untuk menganalisis pengaruh penerapan *feature selection* menggunakan *Pearson Correlation Coefficient (PCC)* pada fitur numerik, *Chi Square* pada fitur kategorikal, serta kombinasi keduanya terhadap performa model *Hybrid ETCXGB* dalam prediksi penyakit jantung.

Dataset yang digunakan berasal dari *UCI Hungarian Heart Disease Dataset*. Tahapan penelitian yang dilakukan meliputi: (1) Pengumpulan Data; (2) *Exploratory Data Analysis (EDA)*; (3) *Preprocessing* melalui penghapusan data duplikat, koreksi nilai yang tidak sesuai dengan domain medis, dan imputasi *missing value* menggunakan median; (4) Seleksi fitur menggunakan *Pearson Correlation Coefficient (PCC)* dan *Chi Square* dengan optimasi parameter melalui *Grid Search*; (5) Implementasi model *ETC*, *XGBoost*, dan *Hybrid ETCXGB*; (6) Evaluasi menggunakan *Stratified Nested Cross Validation* 5×5; (7) Analisis statistik menggunakan uji *Friedman* dan *post hoc Nemenyi*.

Hasil penelitian menunjukkan bahwa skenario *ETCXGB PCC* dan *ETCXGB Combined* menghasilkan rata-rata *Accuracy* tertinggi sebesar 0,8452, diikuti *ETCXGB Chi Square* sebesar 0,8391 dan *ETCXGB Baseline* sebesar 0,8243. Uji *Friedman* menghasilkan nilai statistik 58,7862 dengan *p value* 0,0000016645, yang menunjukkan adanya perbedaan performa yang signifikan untuk 8 dari 18 skenario penelitian. Namun, uji *post hoc Nemenyi* menunjukkan bahwa penerapan *PCC, Chi Square*, maupun kombinasi keduanya pada model *Hybrid ETCXGB* belum menghasilkan perbedaan performa yang signifikan dibandingkan model tanpa *feature selection*. Selain itu, fitur sex, cp, exang, slope, ca, dan thal merupakan fitur yang paling konsisten dipilih pada skenario terbaik. Dengan demikian, penerapan *feature selection* mampu menghasilkan performa yang kompetitif sekaligus mengidentifikasi fitur-fitur klinis yang relevan, meskipun belum terbukti meningkatkan performa model *Hybrid ETCXGB* secara signifikan berdasarkan pengujian statistik.

Kata Kunci: *Chi Square, ETCXGB, Feature Selection,* model *Hybrid, Nested Cross Validation*, *Pearson Correlation Coefficient,* Prediksi Penyakit Jantung, Uji *Friedman*.

ABSTRACT

Heart disease is one of the leading causes of death worldwide, so an accurate and efficient prediction system is needed to support early detection. One approach that is widely used is Machine Learning, including a Hybrid model that combines Extra Trees Classifier (ETC) and Extreme Gradient Boosting (XGBoost). This study aims to analyze the impact of applying feature selection using the Pearson Correlation Coefficient (PCC) on numerical features, Chi Square on categorical features, as well as a combination of both on the performance of the Hybrid ETCXGB model in heart disease prediction.

The dataset used comes from the UCI Hungarian Heart Disease Dataset. The research steps carried out include: (1) Data Collection; (2) Exploratory Data Analysis (EDA); (3) Preprocessing by removing duplicate data, correcting values that don't match medical standards, and imputing missing values using the median; (4) Feature selection using the Pearson Correlation Coefficient (PCC) and Chi Square with parameter optimization through Grid Search; (5) Implementation of ETC, XGBoost, and Hybrid ETCXGB models; (6) Evaluation using 5×5 Stratified Nested Cross Validation; (7) Statistical analysis using the Friedman test and Nemenyi post hoc test.

The results showed that the ETCXGB PCC and ETCXGB Combined scenarios achieved the highest average Accuracy of 0.8452, followed by ETCXGB Chi-Square at 0.8391 and ETCXGB Baseline at 0.8243. The Friedman test produced a statistic of 58.7862 with a p value of 0.0000016645, indicating significant performance differences among 8 of 18 the research scenarios. However, the Nemenyi post hoc test showed that applying PCC, Chi Square, or a combination of both to the Hybrid ETCXGB model did not result in a significant performance difference compared to the model without feature selection. Additionally, the features sex, cp, exang, slope, ca, and thal were the most consistently selected in the best-case scenarios. Therefore, applying feature selection can produce competitive performance while also identifying relevant clinical features, even though it hasn't been proven to significantly improve the performance of the Hybrid ETCXGB model according to statistical tests.

Keywords: Chi Square, ETCXGB, Feature Selection, Hybrid model, Nested Cross Validation, Pearson Correlation Coefficient, Heart Disease Prediction, Friedman Test

DAFTAR ISI

1.1 Latar Belakang 1
1.2 Perumusan Masalah 3
1.3 Tujuan Penelitian 4
1.4 Pembatasan Masalah 4
1.5 Sistematika Penulisan 4
2.1 Tinjauan Penelitian Terdahulu 6
2.2 Landasan Teori 10
2.2.1 Penyakit Jantung 10
2.2.2 *Machine Learning* 11
2.2.3 *Feature selection* 11
2.2.4 *Pearson Correlation Coefficient* 12
2.2.5 *Chi Square* 12
2.2.6 *Hybrid Machine Learning* 13
2.2.7 *Extra Trees Classifier* (*ETC*) 13
2.2.8 *Extreme Gradient Boosting* (*XGBoost*) 13
2.2.9 Uji *Friedman* 14
2.2.10 Uji *Nemenyi* 14
3.1 Perangkat Penelitian 15
3.1.1 Perangkat Keras 15
3.1.2 Pengakat Lunak 15
3.2 Metodologi Penelitian 15
3.2.1 Pengumpulan Data 17
3.2.2 *Exploratory Data Analysis (EDA)* 17
3.2.3 *Preprocessing* 19
3.2.4 Seleksi Fitur 21
3.2.5 Implementasi Model 24
3.2.6 Evaluasi 29
3.2.7 Analisis Statistik 32
3.2.8 Analisis Hasil 34
4.1 Pengumpulan Data 36
4.2 *Exploratory Data Analysis (EDA)* 37

4.3 *Preprocessing*.....44
4.4 Seleksi Fitur 47
4.4.1 *Pearson Correlation Coefficient*.....47
4.4.2 *Chi Square* 51
4.5 Implementasi Model.....55
4.5.1 Validasi Implementasi *Feature selection* pada *Source code* Penelitian Bilal…….58
4.6 Evaluasi 65
4.7 Uji Statistik 78
4.8 Analisis Hasil.....84
5.1 Kesimpulan 90
5.2 Saran.....91

DAFTAR TABEL

Tabel 2. 1 Perbandingan penelitian terdahulu terhadap penelitian yang akan dilakukan....9
Tabel 3. 1 Skenario Penelitian....30
Tabel 4. 1 Dimensi Dataset....37
Tabel 4. 2 Deskripsi Fitur Dataset....38
Tabel 4. 3 Karakteristik Dataset....42
Tabel 4. 4 Nilai Korelasi Fitur Numerik....47
Tabel 4. 5 Nilai Korelasi Seluruh Fitur terhadap Target....49
Tabel 4. 6 Nilai Skor *Chi Square* Fitur Kategorikal....51
Tabel 4. 7 Skor *Chi Square* Seluruh Fitur terhadap Target....53
Tabel 4. 8 Parameter *Grid Search*....56
Tabel 4. 9 Hasil *Nested Cross Validation*....57
Tabel 4. 10 Dataset pada *Source code* Bilal et al.....58
Tabel 4. 11 Hasil Implementasi *Feature selection* pada Dataset dengan Duplikat....59
Tabel 4. 12 Hasil Implementasi *Feature selection* pada Dataset Tanpa Duplikat 61
Tabel 4. 13 Perbandingan Konfigurasi Terbaik pada Kedua Skenario....62
Tabel 4. 14 Perbandingan Penelitian Bilal, Eksperimen Validasi, dan Penelitian Usulan....63
Tabel 4. 15 Konfigurasi Terbaik Setiap Skenario....66
Tabel 4. 16 Hasil *Train* Model....68
Tabel 4. 17 Hasil Evuasi Model....70
Tabel 4. 18 *Accuracy* per*Fold*....78
Tabel 4. 19 Hasil Pengujian *Friedman*....80
Tabel 4. 20 *Ranking Friedman*....80
Tabel 4. 21 Rata-rata *Ranking Friedman*....81
Tabel 4. 22 Skenario yang Berbeda Signifikan....83

DAFTAR GAMBAR

Gambar 3. 1 Metodologi Penelitian 16
Gambar 3. 2 *Flowchart* EDA 18
Gambar 3. 3 *Flowchart Preprocessing* 20
Gambar 3. 4 *Flowchart Pearson Correlation Coefficient* 22
Gambar 3. 5 *Flowchart Chi Square* 23
Gambar 3. 6 *Flowchart Grid Search* 25
Gambar 3. 7 *Flowchart* Implementasi *Feature selection* Pada Penelitian Bilal et al. 28
Gambar 3. 8 *Flowchart* Pengujian Statistik 33
Gambar 4. 1 Repositori *GitHub* Penelitian Bilal et al. 37
Gambar 4. 2 Distribusi Kelas 41
Gambar 4. 3 Cek *Missing value* dan Duplikat 43
Gambar 4. 4 Output Penghapusan Data Duplikat 44
Gambar 4. 5 Perbandingan Distribusi Fitur *Number of Major Vessels* (ca) dan *Thalassaemia* (thal) Sebelum dan Sesudah Koreksi Nilai Tidak Valid 46
Gambar 4. 6 Visualisasi Nilai Korelasi Fitur Numerik 48
Gambar 4. 7 Visualisasi Nilai Korelasi Seluruh Fitur terhadap Target 50
Gambar 4. 8 Visualisasi Nilai Skor *Chi Square* Fitur Kategorikal 52
Gambar 4. 9 Visualisasi Skor *Chi Square* Seluruh Fitur terhadap Target 54
Gambar 4. 10 *Confusion matrix* Konfigurasi Terbaik pada Dataset dengan Duplikat 60
Gambar 4. 11 *Confusion matrix* Konfigurasi Terbaik pada Dataset Tanpa Duplikat 62
Gambar 4. 12 Perbandingan *Train Accuracy* dan *Test Accuracy* Setiap Skenario 72
Gambar 4. 13 *Gap Train Test* Setiap Skenario 73
Gambar 4. 14 *Confusion Matrix* Model *ETC* 75
Gambar 4. 15 *Confusion Matrix* Model *XGB* 76
Gambar 4. 16 *Confusion Matrix* Model *ETCXGB* 77
Gambar 4. 17 Frekuensi Penggunaan Fitur pada Seluruh Skenario Terbaik 86

DAFTAR LAMPIRAN

Lampiran A Hasil Pengujian Statistik A-1

GLOSARIUM

Accuracy	Metrik evaluasi klasifikasi yang mengukur proporsi jumlah prediksi benar terhadap seluruh data yang diuji.
Algorithm	Serangkaian langkah sistematis yang digunakan untuk menyelesaikan suatu permasalahan komputasi.
Angina	Nyeri dada akibat berkurangnya suplai darah dan oksigen ke otot jantung.
Attribute	Karakteristik atau variabel yang merepresentasikan informasi pada dataset.
Bagging	Teknik *ensemble* learning yang membangun beberapa model secara paralel dengan data acak untuk meningkatkan stabilitas prediksi.
Baseline	Skenario awal tanpa penerapan *feature selection* yang digunakan sebagai pembanding terhadap skenario lain.
Bias	Kesalahan sistematis model akibat asumsi yang terlalu sederhana terhadap data.
Binary Classification	Proses klasifikasi data ke dalam dua kelas target, misalnya positif dan negatif.
Boosting	Teknik *ensemble* learning yang membangun model secara bertahap memperbaiki kesalahan model sebelumnya.
Cardiac Disease	Istilah medis untuk penyakit yang menyerang jantung dan sistem kardiovaskular.
Categorical Feature	Fitur yang memiliki nilai berbentuk kategori atau kelas tertentu.
Chi Square	Metode *feature selection* berbasis statistik yang digunakan untuk mengukur hubungan antara fitur kategorikal dan variabel target.
Class Distribution	Persebaran jumlah data pada setiap kelas dalam dataset.
Class Imbalance	Kondisi ketika distribusi jumlah data antar kelas tidak

	seimbang sehingga dapat mempengaruhi performa model klasifikasi.
Classification	Proses prediksi label atau kelas berdasarkan pola data yang dipelajari model *Machine Learning*.
Classifier	Algoritma *Machine Learning* yang digunakan untuk mengelompokkan data ke dalam kelas tertentu.
Cleveland Heart Disease Dataset	Salah satu subset dataset penyakit jantung dari UCI Repository yang sering digunakan pada penelitian *Machine Learning* medis.
Clinical Feature	Variabel medis yang menggambarkan kondisi kesehatan pasien, seperti tekanan darah atau kolesterol.
Combination Feature selection	Pendekatan seleksi fitur yang menggabungkan dua atau lebih metode seleksi fitur.
Confusion Matrix	Tabel evaluasi klasifikasi yang menunjukkan jumlah prediksi benar dan salah berdasarkan kelas aktual dan prediksi model.
Correlation	Hubungan statistik antara dua variabel.
Correlation Coefficient	Nilai numerik yang menunjukkan kekuatan dan arah hubungan antar variabel.
Cross Validation	Teknik evaluasi model dengan membagi data menjadi beberapa subset untuk mengukur kemampuan generalisasi model.
Data cleaning	Proses membersihkan dataset dari data duplikat, *missing value*, noise, dan *outlier* sebelum digunakan dalam pelatihan model.
Data Duplicate	Data yang memiliki nilai atribut identik dengan data lainnya sehingga dapat menyebabkan bias pada model.
Data Mining	Proses menemukan pola atau informasi penting dari sekumpulan data besar.
Data Preprocessing	Tahapan awal pengolahan data sebelum diterapkan pada model *Machine Learning*.
Dataset	Kumpulan data yang digunakan untuk proses analisis, pelatihan, dan evaluasi model *Machine Learning*.

Decision Tree	Algoritma *Machine Learning* berbentuk struktur pohon yang digunakan untuk klasifikasi atau regresi.
Deep Learning	Cabang *Machine Learning* yang menggunakan jaringan saraf tiruan berlapis untuk mempelajari pola kompleks.
Dimensionality Reduction	Proses mengurangi jumlah fitur dalam dataset tanpa kehilangan informasi penting secara signifikan.
Ensemble Learning	Pendekatan *Machine Learning* yang menggabungkan beberapa model untuk meningkatkan performa prediksi.
ETCXGB	Model *Hybrid Machine Learning* yang menggabungkan *Extra Trees Classifier* dan *XGBoost*.
Evaluation Metric	Ukuran yang digunakan untuk mengevaluasi performa model *Machine Learning*.
Exploratory Data Analysis (EDA)	Tahapan analisis awal data untuk memahami distribusi, pola, dan karakteristik dataset.
Extra Trees Classifier (*ETC*)	Algoritma *ensemble* berbasis decision tree yang menggunakan randomisasi tinggi dalam pemilihan split.
F1 Score	Metrik evaluasi klasifikasi yang merupakan harmonic mean dari *Precision* dan *Recall*.
Feature	Variabel atau atribut yang digunakan sebagai input model *Machine Learning*.
Feature Engineering	Proses transformasi dan pengolahan fitur untuk meningkatkan kualitas data dan performa model.
Feature Importance	Tingkat kontribusi suatu fitur terhadap hasil prediksi model.
Feature selection	Proses memilih fitur yang paling relevan untuk meningkatkan efisiensi dan performa model *Machine Learning*.
Filter Method	Metode seleksi fitur berbasis statistik tanpa melibatkan model *Machine Learning* dalam evaluasinya.
Fold	Subset data yang digunakan dalam proses *Cross Validation*.
Friedman Test	Uji statistik non-parametrik untuk membandingkan performa beberapa model atau skenario eksperimen.

Generalization	Kemampuan model untuk memberikan prediksi yang baik pada data baru yang belum pernah dilihat sebelumnya.
Gradient Boosting	Teknik *ensemble* yang membangun model secara iteratif untuk memperbaiki kesalahan model sebelumnya.
Grid Search	Metode pencarian kombinasi *hyperparameter* terbaik secara sistematis berdasarkan ruang parameter tertentu.
Heart Disease	Penyakit yang menyerang jantung dan pembuluh darah.
Heart Disease Prediction	Proses prediksi kemungkinan penyakit jantung menggunakan data medis pasien dan algoritma *Machine Learning*.
Hybrid Model	Model *Machine Learning* yang dibangun dari kombinasi dua atau lebih algoritma.
Hyperparameter	Parameter yang ditentukan sebelum proses pelatihan model dimulai.
Hyperparameter Tuning	Proses optimasi *hyperparameter* untuk memperoleh performa model terbaik.
Imputation	Teknik pengisian *missing value* menggunakan metode tertentu seperti mean atau median.
Inner Cross Validation	Tahapan *Cross Validation* bagian dalam pada *Nested Cross Validation* yang digunakan untuk *tuning* parameter.
K-Fold Cross Validation	Metode *Cross Validation* dengan membagi dataset menjadi K subset untuk proses pelatihan dan pengujian bergantian.
K-Nearest Neighbor (KNN)	Algoritma *Machine Learning* berbasis kedekatan jarak antar data untuk klasifikasi atau regresi.
Label	Variabel target yang menjadi tujuan prediksi model *Machine Learning*.
Linear Relationship	Hubungan antar variabel yang membentuk pola garis lurus.
Machine Learning	Cabang kecerdasan buatan yang memungkinkan sistem belajar dari data tanpa diprogram secara eksplisit.
Mean	Nilai rata-rata dari sekumpulan data numerik.
Median	Nilai tengah dari data yang telah diurutkan.
Missing value	Nilai kosong atau data yang tidak tersedia pada dataset.

Model Training	Proses pelatihan model *Machine Learning* menggunakan data latih.
Multicollinearity	Kondisi ketika beberapa fitur memiliki korelasi sangat tinggi satu sama lain.
NaN (Not a Number)	Representasi nilai kosong atau tidak valid pada data numerik.
Nested Cross Validation	Teknik evaluasi model yang menggunakan *inner* dan *Outer Cross Validation* secara bersamaan.
Noise	Informasi tidak relevan pada data yang dapat menurunkan performa model.
Normalization	Proses penyetaraan skala nilai fitur agar berada dalam rentang tertentu.
Numerical Feature	Fitur berbentuk angka kontinu yang dapat dilakukan operasi matematis.
Outlier	Data yang memiliki nilai jauh berbeda dari distribusi umum data lainnya.
Outer Cross Validation	Tahapan *Cross Validation* bagian luar pada *Nested Cross Validation* untuk evaluasi generalisasi model.
Overfitting	Kondisi ketika model terlalu mempelajari data latih sehingga performanya menurun pada data baru.
Pearson Correlation Coefficient (PCC)	Metode statistik untuk mengukur kekuatan hubungan linear antara dua variabel numerik.
Pipeline	Rangkaian proses otomatis yang menghubungkan *preprocessing*, seleksi fitur, pelatihan, dan evaluasi model.
Post Hoc Test	Uji statistik lanjutan setelah uji utama menunjukkan adanya perbedaan signifikan.
Precision	Metrik evaluasi yang mengukur proporsi prediksi positif yang benar terhadap seluruh prediksi positif.
Prediction	Hasil estimasi model terhadap data baru berdasarkan pola yang telah dipelajari.

Predictive Model	Model yang digunakan untuk memperkirakan hasil atau kelas berdasarkan data historis.
Preprocessing	Tahapan awal pengolahan data sebelum pelatihan model *Machine Learning*.
Random Forest	Algoritma *ensemble* berbasis *decision tree* dengan teknik *bagging* dan *random feature selection* .
Recall	Metrik evaluasi yang mengukur kemampuan model dalam menemukan seluruh data positif.
Redundancy	Kondisi ketika beberapa fitur memiliki informasi yang sama atau berulang.
Root Mean Square Error (RMSE)	Metrik evaluasi untuk mengukur rata-rata kesalahan prediksi berbasis akar kuadrat *error*.
Sampling	Teknik pemilihan sebagian data dari populasi untuk analisis atau pelatihan model.
Scenario	Kombinasi metode atau konfigurasi eksperimen yang diuji dalam penelitian.
Scikit-learn	*Library Python* untuk *Machine Learning* dan analisis data.
Selection Threshold	Nilai ambang batas yang digunakan untuk menentukan fitur yang dipilih.
Standard Deviation	Ukuran statistik yang menunjukkan tingkat penyebaran data terhadap rata-ratanya.
Statistical Test	Metode statistik untuk menguji hipotesis berdasarkan data penelitian.
Stratified K-Fold	Variasi *K-Fold Cross Validation* yang mempertahankan proporsi kelas pada setiap *Fold*.
Supervised Learning	Pendekatan *Machine Learning* yang menggunakan data berlabel untuk pelatihan model.
Target Variable	Variabel keluaran yang diprediksi oleh model *Machine Learning*.
Testing Data	Data yang digunakan untuk mengevaluasi performa model setelah pelatihan.

Threshold	Nilai ambang batas yang digunakan dalam proses seleksi fitur atau pengambilan keputusan model.
Training Data	Data yang digunakan untuk melatih model *Machine Learning*.
UCI Hungarian Heart Disease Dataset	Dataset penyakit jantung dari UCI Repository yang digunakan dalam penelitian ini.
Underfitting	Kondisi ketika model gagal mempelajari pola data secara optimal sehingga performanya rendah.
Unsupervised Learning	Pendekatan *Machine Learning* tanpa menggunakan label target dalam proses pelatihan.
Variance	Tingkat sensitivitas model terhadap perubahan data latih.
Wrapper Method	Metode seleksi fitur yang mengevaluasi subset fitur menggunakan performa model *Machine Learning*.
XGBoost	Algoritma *boosting* berbasis *decision tree* yang dirancang untuk efisiensi komputasi dan performa tinggi.

BAB I
PENDAHULUAN

1.1 Latar Belakang

Penyakit jantung merupakan salah satu penyebab utama kematian di dunia dan masih menjadi tantangan besar dalam bidang kesehatan. Tingginya angka morbiditas dan mortalitas akibat penyakit jantung menunjukkan pentingnya upaya deteksi dini untuk membantu proses diagnosis dan pengambilan keputusan medis secara lebih cepat dan akurat. Namun, proses diagnosis konvensional umumnya memerlukan berbagai pemeriksaan klinis dan laboratorium yang membutuhkan waktu, biaya, serta sumber daya medis yang tidak sedikit. Kondisi tersebut mendorong berkembangnya berbagai pendekatan berbasis teknologi untuk membantu proses identifikasi risiko penyakit jantung secara lebih efisien.

Salah satu pendekatan yang banyak digunakan adalah *Machine Learning* karena kemampuannya dalam mempelajari pola dari data klinis dan menghasilkan prediksi secara otomatis. Penelitian Bharti et al. (2021) menunjukkan bahwa kombinasi *Machine Learning* dan *Deep Learning* mampu menghasilkan performa yang baik dalam prediksi penyakit jantung [1]. Hasil serupa juga dilaporkan oleh Sayadi et al. (2022) yang berhasil membangun model *Machine Learning* untuk mendeteksi *Coronary Artery Disease* menggunakan parameter klinis non-invasif dengan tingkat performa yang tinggi [2]. Temuan-temuan tersebut menunjukkan bahwa *Machine Learning* memiliki potensi besar sebagai sistem pendukung keputusan dalam bidang kesehatan, khususnya untuk prediksi penyakit jantung.

Meskipun demikian, performa model *Machine Learning* tidak hanya ditentukan oleh algoritma klasifikasi yang digunakan, tetapi juga oleh kualitas fitur yang menjadi masukan model. Dataset medis sering kali mengandung fitur yang tidak relevan, redundan, atau memiliki korelasi yang rendah terhadap variabel target sehingga dapat meningkatkan kompleksitas model dan menurunkan kemampuan generalisasi. Alfadli dan Almagrabi (2023) menunjukkan bahwa keterbatasan jumlah dan kualitas fitur dapat memengaruhi performa prediksi pada *UCI Heart Disease Dataset* [3]. Oleh karena itu, diperlukan mekanisme untuk memilih fitur-fitur yang benar-benar berkontribusi terhadap proses klasifikasi.

Feature selection merupakan salah satu teknik yang banyak digunakan untuk mengatasi permasalahan tersebut. Tan et al. (2025) menunjukkan bahwa penerapan *feature selection* yang dioptimalkan mampu meningkatkan performa prediksi penyakit jantung sekaligus mengurangi kompleksitas model [4]. Khan et al. (2024) juga melaporkan bahwa penggunaan *Modified Artificial Bee Colony* (*M-ABC*) untuk seleksi fitur menghasilkan peningkatan akurasi yang signifikan pada prediksi penyakit jantung [5]. Selain itu, kombinasi teknik *balancing* data dan *RFECV feature selection* terbukti mampu meningkatkan performa klasifikasi pada data yang tidak seimbang [6]. Hasil serupa juga ditunjukkan oleh penelitian yang mengembangkan *framework* optimasi *Machine Learning* [7], *adaptive feature selection* [8], *Elastic Net feature selection* [9], *Principal Component Analysis* (*PCA*) [10], serta integrasi *feature selection* dengan *ensemble Deep Learning* [11]. Secara umum, berbagai penelitian tersebut menunjukkan bahwa *feature selection* memiliki peran penting dalam meningkatkan kualitas model prediksi penyakit jantung.

Di sisi lain, perkembangan penelitian juga menunjukkan bahwa model *Hybrid* dan *ensemble* cenderung menghasilkan performa yang lebih baik dibandingkan model tunggal. Majumder et al. (2023) menunjukkan bahwa penggabungan beberapa classifier dalam sebuah *Hybrid ensemble* mampu meningkatkan akurasi prediksi penyakit jantung [12]. Hasil yang lebih mutakhir dilaporkan oleh Bilal et al. (2024) melalui pengembangan model *Hybrid ETCXGB* yang mengombinasikan *Extra Trees Classifier* (*ETC*) dan *Extreme Gradient Boosting* (*XGBoost*), serta menghasilkan performa yang sangat baik dalam prediksi penyakit jantung [13]. Temuan tersebut menunjukkan bahwa kombinasi algoritma *ensemble* dapat menjadi pendekatan yang menjanjikan untuk meningkatkan akurasi klasifikasi.

Walaupun penelitian mengenai *feature selection* dan model *Hybrid* telah berkembang pesat, sebagian besar penelitian terdahulu lebih berfokus pada pengembangan metode seleksi fitur atau peningkatan arsitektur model secara terpisah. Masih terdapat keterbatasan penelitian yang secara khusus menganalisis pengaruh metode *feature selection* berdasarkan karakteristik tipe data terhadap performa model *Hybrid*. Padahal, dataset penyakit jantung umumnya terdiri atas kombinasi fitur numerik dan kategorikal yang memiliki karakteristik berbeda. *Pearson Correlation Coefficient (PCC)* banyak digunakan untuk mengukur hubungan linear pada fitur

numerik, sedangkan *Chi Square* baik dalam mengukur relevansi fitur kategorikal terhadap variabel target. Penelitian Sarra et al. (2024) menunjukkan bahwa penerapan *Chi Square Feature selection* mampu meningkatkan performa model *ensemble* pada prediksi penyakit jantung. Namun demikian, penelitian yang menganalisis pengaruh penerapan *PCC* pada fitur numerik dan *Chi Square* pada fitur kategorikal secara bersamaan terhadap model *Hybrid ETCXGB* masih terbatas [14].

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk menganalisis pengaruh penerapan *feature selection* menggunakan *Pearson Correlation Coefficient (PCC)* pada fitur numerik dan *Chi Square* pada fitur kategorikal terhadap performa model *Hybrid ETCXGB* yang dikembangkan berdasarkan penelitian Bilal et al. (2023). Analisis dilakukan menggunakan *UCI Hungarian Heart Disease* Dataset untuk mengevaluasi sejauh mana penerapan *feature selection* berbasis tipe data mampu meningkatkan performa prediksi penyakit jantung. Hasil penelitian ini diharapkan dapat memberikan kontribusi akademik berupa pemahaman yang lebih mendalam mengenai pengaruh *feature selection* berbasis tipe data terhadap model *Hybrid Machine Learning*, sekaligus memberikan kontribusi praktis dalam pengembangan sistem prediksi penyakit jantung yang lebih efisien dan akurat.

1.2 Perumusan Masalah

Berdasarkan uraian dari latar belakang tersebut, penelitian ini dirancang dengan perumusan masalah sebagai berikut:

1. Apakah penerapan metode *feature selection Pearson Correlation Coefficient (PCC)* untuk fitur numerik berpengaruh terhadap performa model *Hybrid Machine Learning ETCXGB* dalam prediksi penyakit jantung?
2. Apakah penerapan metode *feature selection Chi Square* untuk fitur kategorikal berpengaruh terhadap performa model *Hybrid Machine Learning ETCXGB* dalam prediksi penyakit jantung?
3. Apakah penerapan kombinasi metode *PCC* dan *Chi Square* menghasilkan peningkatan performa model *Hybrid Machine Learning ETCXGB* dibandingkan dengan penerapan salah satu metode saja?

1.3 Tujuan Penelitian

Berdasarkan perumusan masalah yang telah dijelaskan, penelitian ini memiliki tujuan sebagai berikut:

1. Untuk mengetahui pengaruh penerapan metode *PCC* terhadap hasil evaluasi model *Hybrid Machine Learning ETCXGB* dalam prediksi penyakit jantung.
2. Untuk mengetahui pengaruh penerapan metode *Chi Square* terhadap hasil evaluasi model *Hybrid Machine Learning ETCXGB* dalam prediksi penyakit jantung.
3. Untuk mengetahui apakah kombinasi metode *PCC* dan *Chi Square* menghasilkan peningkatan hasil evaluasi model *Hybrid Machine Learning ETCXGB* dibandingkan dengan penggunaan salah satu metode secara terpisah.

1.4 Pembatasan Masalah

Adapun pembatasan masalah yang ditetapkan dalam penelitian ini mencakup sebagai berikut:

1. Penelitian ini menggunakan dataset *UCI Hungarian Heart Disease* yang ditelah digunakan pada penelitian yang menjadi referensi penelitian ini
2. Proses *feature selection* dibatasi pada dua metode yaitu *Pearson Correlation Coefficient (PCC)* untuk fitur numerik dan *Chi Square* untuk fitur kategorikal.
3. Model prediksi yang digunakan adalah Hybrid *Extra Trees Classifier* (*ETC*) dan *Extreme Gradient Boosting* (*XGBoost*).
4. Evaluasi performa model dibatasi pada metrik *Accuracy*, *Precision*, *Recall*, dan *F1 Score*.

1.5 Sistematika Penulisan

Sistematika penulisan skripsi ini disusun agar memudahkan pembahasan penelitian secara terstruktur. Adapun susunannya adalah sebagai berikut:

BAB I : PENDAHULUAN Bab ini memuat latar belakang penelitian, rumusan masalah, tujuan dan manfaat penelitian, batasan masalah, serta sistematika penulisan.

BAB II : TINJAUAN PUSTAKA Bab ini membahas landasan teori yang mendukung penelitian, meliputi *Extra Trees Classifier* (*ETC*), *Extreme Gradient Boosting* (*XGBoost*), model *Hybrid*, metode *feature selection* (*Pearson Correlation*

Coefficient dan *Chi Square*), serta uji statistik *Friedman*. Selain itu, disajikan pula kajian penelitian terdahulu yang relevan.

BAB III : METODOLOGI PENELITIAN Bab ini menjelaskan desain eksperimen penelitian, sumber dan karakteristik dataset, tahapan *preprocessing*, penerapan metode *feature selection* (*PCC* untuk fitur numerik dan *Chi Square* untuk fitur kategorikal), implementasi model *Hybrid ETCXGB*, skenario pengujian, serta metode evaluasi kinerja model menggunakan *Nested Cross Validation* dan uji statistik *Friedman*.

BAB IV : HASIL DAN ANALISIS Bab ini menyajikan hasil eksperimen dari setiap skenario *feature selection* , analisis perbandingan performa model berdasarkan metrik evaluasi, interpretasi hasil uji statistik, serta pembahasan implikasi hasil terhadap efektivitas penerapan *feature selection* pada model *Hybrid ETCXGB*.

BAB V : PENUTUP Bab ini berisi kesimpulan yang menjawab rumusan masalah berdasarkan hasil penelitian yang diperoleh, serta saran dan rekomendasi untuk pengembangan penelitian selanjutnya.