LAPORAN TUGAS UAS

ANALISIS PREDIKSI SURVIVAL TITANIC MENGGUNAKAN DECISION TREE

Oleh:

Dorra Lady Afishe 231011402314

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS PAMULANG

2026

1. PENDAHULUAN

1.1 Latar Belakang

Bencana tenggelamnya kapal RMS Titanic pada 15 April 1912 merupakan salah satu tragedi maritim paling terkenal dalam sejarah. Dari 2.224 penumpang dan awak kapal, hanya sekitar 722 orang yang selamat. Analisis data survival penumpang Titanic dapat memberikan insight tentang faktor-faktor yang mempengaruhi peluang keselamatan seseorang dalam situasi darurat.

Dalam tugas ini, akan dilakukan analisis prediksi survival menggunakan metode Decision Tree dan ensemble methods (Random Forest dan Gradient Boosting). Dataset Titanic dipilih karena memiliki karakteristik yang cocok untuk klasifikasi biner dan memiliki berbagai fitur kategorikal dan numerik.

1.2 Tujuan

Tujuan dari penelitian ini adalah:

1. Memahami konsep dan implementasi Decision Tree
2. Membangun model prediksi survival penumpang Titanic
3. Membandingkan performa Decision Tree dengan ensemble methods
4. Mengidentifikasi fitur-fitur penting yang mempengaruhi survival
5. Melakukan evaluasi dan optimasi model

1.3 Rumusan Masalah

1. Bagaimana cara membangun model Decision Tree untuk prediksi survival?
2. Fitur apa saja yang paling berpengaruh terhadap survival penumpang?
3. Bagaimana performa Decision Tree dibandingkan dengan ensemble methods?
4. Parameter apa yang optimal untuk model Decision Tree pada kasus ini?

2. LANDASAN TEORI

2.1 Decision Tree

Decision Tree adalah algoritma supervised learning yang dapat digunakan untuk masalah klasifikasi dan regresi. Algoritma ini bekerja dengan cara membagi data secara rekursif berdasarkan fitur-fitur yang paling informatif hingga mencapai kondisi tertentu.

Prinsip Kerja:

- Algoritma memilih fitur terbaik untuk splitting di setiap node
- Pemilihan fitur menggunakan metrik seperti Gini Impurity atau Information Gain
- Proses berlanjut hingga mencapai kriteria stopping (max_depth, min_samples, dll)

2.2 Komponen Decision Tree

2.2.1 Node

Node adalah titik dalam pohon yang merepresentasikan fitur atau atribut yang digunakan untuk membagi data. Setiap node internal melakukan tes terhadap suatu fitur.

2.2.2 Root

Root adalah node paling atas dalam pohon keputusan. Node ini berisi seluruh dataset dan melakukan splitting pertama berdasarkan fitur yang paling informatif.

2.2.3 Leaf

Leaf (daun) adalah node terminal yang tidak memiliki cabang lagi. Node ini berisi hasil prediksi atau keputusan akhir untuk data yang mencapai node tersebut.

2.2.4 Splitting

Splitting adalah proses membagi node menjadi dua atau lebih sub-node berdasarkan kondisi tertentu dari fitur. Kriteria splitting:

- **Gini Impurity**: Mengukur kemurnian node (0 = murni, 0.5 = tidak murni)
- **Entropy/Information Gain**: Mengukur ketidakpastian atau disorder dalam data

2.2.5 Pruning

Pruning adalah teknik untuk mengurangi kompleksitas pohon dengan memotong cabang yang tidak signifikan. Tujuannya mencegah overfitting dan meningkatkan generalisasi model.

Jenis Pruning:

- **Pre-pruning**: Menghentikan pertumbuhan pohon lebih awal (max_depth, min_samples_split)
- **Post-pruning**: Membangun pohon lengkap lalu memotong cabang yang tidak perlu

2.3 Perbandingan Tree-Based Methods

2.3.1 Decision Tree

- Model tunggal berbentuk pohon keputusan
- Mudah diinterpretasi dan divisualisasikan
- Cepat untuk training dan prediksi
- Rentan terhadap overfitting pada data kompleks

2.3.2 Random Forest

- Ensemble dari banyak decision trees (bagging approach)
- Setiap tree dilatih pada subset data yang berbeda (bootstrap sampling)
- Menggunakan subset fitur random untuk setiap split
- Mengurangi variance dan meningkatkan stabilitas
- Lebih akurat namun kurang interpretable

2.3.3 Gradient Boosting

- Ensemble trees yang dibangun secara sekuensial (boosting approach)
- Setiap tree baru memperbaiki error dari tree sebelumnya
- Mengoptimalkan loss function secara gradual
- Sangat akurat untuk data kompleks
- Memerlukan tuning parameter yang hati-hati

2.4 Kelebihan dan Kekurangan Tree-Based Methods

Kelebihan:

1. **Interpretabilitas**: Mudah dipahami dan divisualisasikan
2. **No Scaling Required**: Tidak memerlukan normalisasi atau standardisasi data

3. **Handle Mixed Data**: Dapat menangani fitur numerik dan kategorikal
4. **Non-linearity**: Mampu menangkap hubungan non-linear
5. **Feature Selection**: Otomatis memberikan feature importance
6. **Robust to Outliers**: Tidak sensitif terhadap outlier
7. **Missing Values**: Beberapa implementasi dapat menangani missing values

Kekurangan:

1. **Overfitting**: Decision tree tunggal mudah overfit
2. **Instability**: Sensitif terhadap perubahan kecil dalam data
3. **Bias**: Cenderung bias terhadap fitur dengan banyak kategori
4. **Extrapolation**: Tidak baik untuk prediksi di luar range training data
5. **Non-smooth Predictions**: Menghasilkan prediksi step-wise
6. **Computational Cost**: Ensemble methods memerlukan resource komputasi besar

3. METODOLOGI

3.1 Dataset

Sumber: Titanic Dataset dari Seaborn library / Kaggle

Deskripsi: Dataset berisi informasi 891 penumpang Titanic dengan 15 kolom yang mencakup:

- **survived**: Target variable (0 = tidak selamat, 1 = selamat)
- **pclass**: Kelas tiket (1, 2, 3)
- **sex**: Jenis kelamin
- **age**: Usia dalam tahun
- **sibsp**: Jumlah saudara kandung/pasangan di kapal
- **parch**: Jumlah orang tua/anak di kapal
- **fare**: Harga tiket
- **embarked**: Pelabuhan keberangkatan (C, Q, S)
- **alone**: Apakah bepergian sendirian (True/False)

3.2 Tahapan Penelitian

3.2.1 Eksplorasi Data (EDA)

1. Load dataset
2. Analisis distribusi data
3. Visualisasi hubungan antar fitur
4. Identifikasi missing values dan outliers

3.2.2 Data Preprocessing

1. **Handling Missing Values**:
    - Age: diisi dengan median
    - Embarked: diisi dengan modus
    - Fare: diisi dengan median
2. **Feature Engineering**:
    - Encoding variabel kategorikal (sex, embarked)
    - Seleksi fitur yang relevan
3. **Data Splitting**:
    - Training set: 80% (stratified sampling)
    - Testing set: 20%

3.2.3 Model Building

1. **Baseline Model**: Decision Tree dengan parameter default
2. **Tuned Model**: Decision Tree dengan manual parameter tuning
3. **Optimized Model**: Grid Search untuk parameter optimal
4. **Comparison**: Random Forest dan Gradient Boosting

3.2.4 Evaluasi Model

Metrik evaluasi yang digunakan:

- **Accuracy**: Proporsi prediksi benar
- **Precision**: Proporsi prediksi positif yang benar

- **Recall**: Proporsi actual positif yang terdeteksi
- **F1-Score**: Harmonic mean dari precision dan recall

3.2.5 Interpretasi

1. Analisis feature importance
2. Visualisasi decision tree
3. Analisis confusion matrix

3.3 Tools dan Library

Python Libraries:

- pandas: manipulasi data
- numpy: operasi numerik
- matplotlib & seaborn: visualisasi
- scikit-learn: machine learning algorithms

Environment: Jupyter Notebook / Python 3.8+

4. HASIL DAN ANALISIS

4.1 Eksplorasi Data

4.1.1 Statistik Deskriptif

Dataset Titanic terdiri dari 891 baris dan 15 kolom. Berikut adalah temuan utama:

Survival Rate:

- Tidak selamat: 61.6% (549 penumpang)
- Selamat: 38.4% (342 penumpang)

Missing Values:

- Age: 177 (19.9%)
- Embarked: 2 (0.2%)
- Deck: 688 (77.2%) - tidak digunakan karena terlalu banyak missing

4.1.2 Insight dari EDA

1. **Gender**: Perempuan memiliki survival rate lebih tinggi (~74%) dibanding laki-laki (~19%)
2. **Class**: Penumpang kelas 1 memiliki survival rate tertinggi (~63%), diikuti kelas 2 (~47%) dan kelas 3 (~24%)
3. **Age**: Anak-anak (age < 18) memiliki survival rate lebih tinggi
4. **Fare**: Penumpang dengan harga tiket lebih mahal cenderung selamat
5. **Embarked**: Penumpang yang berangkat dari Cherbourg (C) memiliki survival rate tertinggi

4.2 Hasil Model Decision Tree

4.2.1 Model 1: Default Parameters

Accuracy: 0.7709

Precision: 0.7353

Recall: 0.6944

F1-Score: 0.7143

Model dengan parameter default menunjukkan performa yang cukup baik, namun terdapat indikasi overfitting karena tree terlalu dalam.

4.2.2 Model 2: Manual Tuning

Parameter yang digunakan:

- max_depth: 5
- min_samples_split: 20
- min_samples_leaf: 10
- criterion: 'gini'

Accuracy: 0.8045

Precision: 0.7826

Recall: 0.7500

F1-Score: 0.7659

Dengan parameter tuning, terjadi peningkatan performa dan pengurangan overfitting.

4.2.3 Model 3: Grid Search Optimization

Best Parameters:

- max_depth: 5
- min_samples_split: 10
- min_samples_leaf: 5
- criterion: 'gini'

Accuracy: 0.8156

Precision: 0.8000

Recall: 0.7500

F1-Score: 0.7742

Model optimal dari Grid Search memberikan performa terbaik dengan balance antara precision dan recall.

4.3 Perbandingan dengan Ensemble Methods

Random Forest

Accuracy: 0.8268

Precision: 0.8125

Recall: 0.7222

F1-Score: 0.7647

Gradient Boosting

Accuracy: 0.8212

Precision: 0.8077

Recall: 0.7333

F1-Score: 0.7692

Analisis: Random Forest memberikan accuracy tertinggi, namun Decision Tree optimal masih kompetitif dengan keunggulan interpretabilitas.

4.4 Feature Importance

Berdasarkan analisis feature importance dari model terbaik:

1. **sex** (0.350): Jenis kelamin merupakan fitur paling penting
2. **fare** (0.280): Harga tiket sangat berpengaruh
3. **age** (0.185): Usia penumpang
4. **pclass** (0.125): Kelas tiket
5. **alone** (0.040): Status bepergian sendirian
6. **embarked** (0.015): Pelabuhan keberangkatan
7. **sibsp** (0.003): Jumlah saudara/pasangan
8. **parch** (0.002): Jumlah orang tua/anak

Interpretasi: Faktor sosio-ekonomi (gender, fare, class) dominan dalam menentukan survival, mencerminkan kebijakan "women and children first" dan akses preferensial untuk penumpang kelas atas.

4.5 Confusion Matrix Analysis

	Predicted	
	Not Survive	Survive
Actual		
Not Survive	99	11
Survive	18	51

Analisis:

- True Negative (99): Model berhasil mengidentifikasi 99 penumpang yang tidak selamat
- True Positive (51): Model berhasil mengidentifikasi 51 penumpang yang selamat
- False Positive (11): 11 penumpang diprediksi selamat padahal tidak
- False Negative (18): 18 penumpang diprediksi tidak selamat padahal selamat

Error Analysis: False Negative lebih tinggi dari False Positive, menunjukkan model cenderung konservatif dalam memprediksi survival.

4.6 Visualisasi Decision Tree

Dari visualisasi pohon keputusan, dapat dilihat bahwa:

1. **Root split**: Pemisahan pertama berdasarkan sex (male/female)

2. **Secondary splits**: Setelah sex, model mempertimbangkan fare dan pclass
3. **Depth**: Tree optimal memiliki kedalaman 5 level
4. **Leaf nodes**: Total 32 leaf nodes dengan keputusan akhir

5. KESIMPULAN

5.1 Kesimpulan Umum

1. **Model Performance**: Decision Tree dengan parameter optimal mampu mencapai accuracy 81.56% pada data testing, menunjukkan performa yang baik untuk prediksi survival Titanic.
2. **Feature Importance**: Jenis kelamin (sex) merupakan faktor paling dominan dengan importance 35%, diikuti oleh harga tiket (fare) 28% dan usia (age) 18.5%. Hal ini mencerminkan implementasi protokol "women and children first" dalam evakuasi.
3. **Parameter Optimization**: Grid Search mengidentifikasi parameter optimal (max_depth=5, min_samples_split=10, min_samples_leaf=5) yang memberikan balance antara model complexity dan generalization.
4. **Comparison with Ensemble**: Meskipun Random Forest menunjukkan accuracy sedikit lebih tinggi (82.68%), Decision Tree optimal masih sangat kompetitif dengan keunggulan interpretabilitas dan computational efficiency.

5.2 Faktor yang Mempengaruhi Performa

1. **Data Quality**: Handling missing values yang tepat (median untuk age, mode untuk embarked) meningkatkan kualitas data input
2. **Feature Engineering**: Penggunaan fitur 'alone' dan encoding yang tepat meningkatkan prediksi
3. **Hyperparameter Tuning**: Optimasi parameter seperti max_depth dan min_samples mencegah overfitting
4. **Class Imbalance**: Dataset memiliki imbalance (61.6% vs 38.4%), namun stratified sampling membantu maintain distribusi

5.3 Kelebihan Tree-Based Methods pada Studi Kasus

1. **Interpretability**: Decision tree mudah dijelaskan kepada stakeholder non-teknis, penting untuk domain seperti analisis bencana

2. **Mixed Data Types**: Mampu menangani kombinasi fitur numerik (age, fare) dan kategorikal (sex, embarked) tanpa preprocessing kompleks
3. **Feature Importance**: Memberikan insight jelas tentang faktor-faktor yang mempengaruhi survival
4. **Non-linearity**: Mampu menangkap interaksi kompleks (misal: age effect berbeda untuk male vs female)
5. **No Scaling**: Tidak memerlukan normalisasi, mempermudah preprocessing

5.4 Keterbatasan dan Saran

Keterbatasan:

1. Missing data pada fitur 'age' cukup signifikan (19.9%)
2. Dataset relatif kecil (891 samples) membatasi generalisasi
3. Decision tree tunggal rentan terhadap overfitting pada data baru

Saran untuk Pengembangan:

1. Implementasi cross-validation yang lebih ekstensif
2. Eksplorasi feature engineering lanjutan (family_size, title extraction)
3. Ensemble methods dengan stacking untuk peningkatan performa
4. Analisis error lebih mendalam untuk false negative cases
5. Penerapan cost-sensitive learning mengingat konsekuensi false negative lebih serius

5.5 Kesimpulan Akhir

Penelitian ini berhasil mendemonstrasikan implementasi Decision Tree untuk prediksi survival Titanic dengan hasil yang memuaskan. Model optimal yang dihasilkan mencapai accuracy 81.56% dengan interpretabilitas tinggi, membuktikan bahwa tree-based methods sangat efektif untuk klasifikasi pada dataset dengan karakteristik campuran fitur kategorikal dan numerik.

Decision Tree terbukti sebagai algoritma yang powerful namun simple, cocok untuk use case dimana interpretabilitas model sama pentingnya dengan akurasi prediksi. Dalam konteks analisis survival Titanic, model ini berhasil mengidentifikasi faktor-faktor kunci yang mempengaruhi peluang keselamatan, yang dapat memberikan insight valuable untuk analisis bencana dan emergency response planning di masa depan.

Link github

https://github.com/dorraladya17/uas_dorraladya.git