Analisa Faktor Keselamatan Dari Bencana Titanic

Jeremiah Dylan Julianto  
School of Computer Science  
Bina Nusantara UniversityTangerang, Indonesia  
Daniel Alexander  
School of Computer Science  
Bina Nusantara UniversityTangerang, Indonesia  
Thomas Amarta Gunawisesa  
School of Computer Science  
*Bina Nusantara University*Tangerang,Indonesia  
Divaldy Putra  
School of Computer Science  
*Bina Nusantara University*  
Tangerang,Indonesia

Christopher Nathanael  
School of Computer Science  
Bina Nusantara UniversityTangerang, Indonesia

Abstract - Titanic merupakan tragedi maritim yang besar dan cukup terkenal. Tragedi ini terjadi pada 15 April 1912 dan memakan lebih dari 1500 korban jiwa baik dari crew maupun penumpang yang ada pada kapal. Tragedi ini dikatakan terjadi karena kapal RMS Titanic menabrak sebuah gunung es yang merusak hulu kapal yang membuat sisi kanan kapal kemasukan air hingga akhirnya tenggelam. Walaupun banyak terdapat korban jiwa, akan tetapi tetap terdapat yang berhasil selamat dari tragedi tersebut. Maka dari itu, kami ingin mempelajari dan menganalisa potensi survivability rate dan faktor apa saja yang mempengaruhi potensi tersebut.

Keywords— titanic, machine learning, data analysis

# Pendahuluan

Dalam menganalisa Tragedi Titanic ada banyak hal yang bisa di analisa. Oleh karena itu, kami ingin menganalisa variabel apa saja yang mempengaruhi faktor keselamatan dari penumpang melalui dataset ini. Kami juga ingin mempelajari bagaimana memprediksi jumlah orang yang selamat pada tragedi tersebut.

Pada kasus titanic ini kami ingin mengetahui **classification** dari variabel survived. Oleh karena itu kami menggunakan supervised machine learning algorithm. Random Forest Classifier adalah algoritma yang akan kami gunakan pada percobaanini.

Pada projek ini kami akan berfokus melakukan visualisasi data dan Analisa data dengan metode *basic analysis* dan *prediction analysis(Machine Learning)* untuk mengetahui factor apa saja yang mempengaruhi variable *Survived* serta performa *Random Forest* dalam permasalah dataset Titanic*.*

# Metodologi

## Import Library

Ada beberapa library yang kami gunakan pada projek ini untuk membaca file, kalkulasi matematika, normalisasi data, dan masih banyak lagi. Untuk detailnya kami tampilkan pada Table 1

|  |  |
| --- | --- |
| Library | Function |
| Numpy | Mathematical Function |
| Pandas | File Reading |
| Sklearn.Preprocessing.LabelEncoder | Data Normalization |
| Sklearn.model\_selection.Train\_test\_split | Split arrays or matrices into random train and test subsets. |
| Matplotlib.Gridspec | Maintain aesthetics of the data visualisation |
| matplotlib.pyplot | Provides an implicit, MATLAB-like, way of plotting |
| seaborn | Data visualization |
| Sklearn.ensemble.RandomForestClassifier | Machine Learning |
| Sklearn.model\_selection   * Learning\_curve * ShuffleSplit | * Estimate ML model performance * Split data into train and test dataset |

## Gambaran Data dan Eksplorasi Data

Dataset yang kami gunakan yaitu dataset seluruh penumpang yang menaiki kapal Titanic yang diambil dari kaggle. Dataset tersebut berbentuk *comma separated file* (.csv). Pada file column, terdapat data-data penumpang seperti PassengerID, Nama, Sex, Age, dll. Data type dari data-data tersebut adalah object, float, atau integer dengan detail yang ada pada Table 1. Dataset yang kami pakai berjumlah 891.

Table

Description automatically generated

Table 1 . Informasi Detail Dataset

A screenshot of a computer

Description automatically generated with medium confidence

Figure 1. Tampilan Isi Data

A picture containing text, window, scoreboard

Description automatically generated

Figure 2. Tampilan pd.describe()

Kami menjalankan function *pd.describe()* dan kami menemukan anomaly data dimana terdapat min Fare = 0. Setelah kami telusuri lebih lanjut terdapat bahwa ditemukan beberapa penumpang dengan *Ticket* yang sama dan memiliki nilai *Fare =0.* Setelah kami telusuri lebih lanjut [www.encylopedia-titanica.org](http://www.encylopedia-titanica.org/) terdapat bahwa beberapa awak kapal dari Titanic memberikan ticket gratis ke pada beberapa penumpang oleh karena itu ditemukan lah penumpang dengan *Ticket* yang sama dan *Fare* =0 pada Figure 3.

A picture containing text

Description automatically generated

Figure 3. Tampilan Ticket dan Fare

## Data Cleansing

Setelah kami menganalisa data kami lebih lanjut kami menemukan bahwa ada beberapa data yang memiliki nilai NULL Figure 4.

A screenshot of a computer

Description automatically generated with low confidence

Figure 4. NULL Data

Untuk mengatasi masalah tersebut kami memutuskan solusi berikut:

1. Untuk data Age kami akan menggunakan mean Ticket
2. Untuk data Cabin kami akan menghapus kolom tersebut karena memiliki banyak nilai NULL yang dapat mempengaruhi hasil akhir.
3. Untuk data Embarked kami akan menggunakan algoritma label encoding untuk mengisi kekosongan tersebut.

Kami melakukan drop pada beberapa variable yang tidak berhubungan dengan tujuan projek kami.

1. PassangerId
2. Ticket
3. Name
4. Fare

## Data Normalization

Kami melakukan proses *grouping* untuk variable *age* dikarenakan memiliki rentan umur yang beragam dan dapat mempermudah proses train dari machine learning itu sendiri. Untuk grouping kami menggunakan referensi dari [5] dengan detail pada Table 2

|  |  |
| --- | --- |
| Age | New Group |
| < 12 | 0 |
| 13- 18 | 1 |
| 19 – 59 | 2 |
| 60 | 3 |

Table 2. Pengkategorian Umur

## Machine Learning

Kami akan menggunakan scikit-learn library untuk mengimplementasikan RandomForestClassifier Algorithm kami. Dengan konfigurasi *hyperparameter* *default* Table 3*:*

|  |  |
| --- | --- |
| n\_estimator | 100 |
| max\_depth | None |
| min\_leaf\_node:1 | 1 |
| max\_leaf\_nodes:None | None |

Table 3. *Hyperparamater*

## Features

A screenshot of a computer

Description automatically generated with low confidence

Figure 5. Heatmap

Dari hasil heatmap pada Figure 5 kami menambahkan dua features baru yaitu:

* *Positive Correlation (Parch, FamilySize)*
* *Negative Correlation (Pclass, Sex, Age, SibSp, Embarked, IsAlone)*

## Data Splitting

Kami membagi dengan metode pareto law yang terbukti

memberikan hasil lebih optimal disbanding proporsi partisi yang lain [1,2]. Kami akan membaginya sesuai dengan features yang ada *All, Positive,* dan *Negative*.

Target = Survived

Features= dfTrain.drop(‘Survived’,axis=1)

All Features

Text

Description automatically generated

Positive Features

Text

Description automatically generated

Negative Features

Text

Description automatically generated

## Training

Training kami lakukan dengan library fitur

*RandomForestClassifier* , kami membaginya kedalam tiga train berbeda sesuai dengan hasil *features* yang kami buat. Berikut kami cantukan hasil training dari masing-masing *features.*

All Features



Positive Features

A screenshot of a computer

Description automatically generated with medium confidence

Negative Features

Text

Description automatically generated

# Hasil

## Basic Analysis

Kami melakukan Basic Analysis dengan menggunakan library seasborn dimana untuk mengambil konklusi dari data yang ada kami menggunakan *seasborn.barplot()*. Hal ini disebabkan karena data kami berupa data *survived* meruapakan data klasifikasi bernilai 1/0 sehingga kami ingin melakukan perbandingan ratio secara menyeluruh dan kami dimana menggunakan *mean* sebagai pembanding ratio tersebut.

Chart, bar chart

Description automatically generated

Figure 6. Hasil Basic Analysis

Dari hasil Analisa Figure 6 kami mendapatkan kesimpulan yaitu

|  |  |
| --- | --- |
| Sex | Female (0.74) memiliki ratio keselamatan lebih tinggi disbanding pria (0.18) |
| Pclass | Penumpang pada Pclass 1 memiliki ratio keselamatan tertinggi yaitu 0.63 |
| SibSp | Penumpang dengan jumlah SibSp 1 atau 2 memiliki peluang keselamatan tinggi |
| Parch | Penumpang dengan jumlah 1, 2, dan 3 memiliki peluang keselamatan tinggi. |
| Age | Penumpang yang berada di age group 0 (<12) memiliki peluang selamat tertinggi. |
| Embarked | Penumpang yang pergi dari Perancis memiliki peluang keselamatan tertinggi |

Table 3. Hasil Basic Analysis

Dari kesimpulan diatas kami menemukan bahwa ada beberapa variable yang menarik untuk diinvestigasi lebih lanjut yang akan kami sampaikan pada poin B.

## Deep analysis

Dari hasil Basic analysis kami menemukan bahwa ada

variable SibSp, Parch, Embarked, dan Sex memiliki keunikan dengan detail yang ada pada Tabel 4.

|  |  |
| --- | --- |
| Variabel | Keunikan |
| SibSp dan Parch | SibSp memiliki relasi dengan Parch. Ketikajumlah SibSp = 0 dan Parch =0 memiliki probabilitas keselamtan yang sama namun disini kita melihat ada perbedaan cukup siginifikan dimana ketika SibSp = 3 dan Parch =3 memiliki perbedaan ratio yang tinggi.  Sehingga kami memutuskan untuk membuat kolom baru yaitu *Family Size* dan *IsAlone* untuk mengetahui lebih jelas relasi yang ada. |
| Embarked | Dari data ini kita tahu bahwa Keberangkatan dari Peranci memiliki peluang keselamtan tinggi namun kami ingin mengetahui apa factor yang menyebabkan terjadinya hal tersebut dan korelasinya dengan variable lain. |
| Sex | Dari variabl ini *Female* memiliki probabilitas keselamatan yang sangat tinggi hingga >70% dimana *Male*  hanya memiliki angka <20% sehingga kami ingin mengetahui apa yang membuat variable *Female* ini sangat spesial |

Table 4. Alasan Deep Analysis

1. SibSp dan Parch

Family size dan IsAlone

Kami menggabungkan variable *SibSp* dan *Parch* kedalam satu variable *FamilySize.*

*Text

Description automatically generated*

Dengan tujuan ingin mengkaji lebih dan mendapatkan gambaran terkait penumpang yang berpegian tidak sendiri.

Chart, bar chart

Description automatically generated

Figure 7. Ratio Family Size Terhadap Survied

Text

Description automatically generatedPada Figure 7 kita dapat membuat kesimpulan baru dimana kita dapat melihat dengan jelas bahwa penumpang dengan *FamilySize 2,3,4* memiliki peluang untuk selamat tinggi dibanding yang lain. Untuk mebuktikan statement bahwa penumpang yang tidak sendiri memiliki peluang keselematan tinggi kami juga membuat variable IsAlone.

Chart, bar chart

Description automatically generated

Figure 8. Ratio Size IsAlone Terhadap Survived

Dari Figure 8 kita mendapat kepastian bahwa penumpang yang tidak sendiri memiliki peluang selamat lebih tinggi.

1. Embarked

Kami melakukan Analisa dengan menggunakan *seasborn.countplot()* dengan tujuan ingin membandingkan antar variable yang dihubungkan dengan Table 3.

A picture containing bar chart

Description automatically generated

Figure 9. Countplot Embarked (C)

Dari hasil figure 9 diatas dan hasil berdasarkan Table 3 kita dapat mendapat kesimpulan dengan detail pada Table 5.

|  |
| --- |
| 1. Tidak didasari faktor Family Size |
| 2. Tidak didasari oleh Umur |
| 3. Memiliki jumlah jumlah Female lebih dominan dibanding Male |
| 4. Memiliki jumlah penumpang kelas 1 yang lebih dominan dibanding Pclass lainnya |

Table 5. Hasil Deep Analysis Embarked

1. Sex

Chart

Description automatically generated with medium confidence

Setelah melakukan Analisa dengan melakukan visualisasi menggunakan *seasborn.countplot()*. Kami mendapatkan kesimpulan bahwa variable *Sex* memang mempengaruhi ratio keselamatan tertinggi dibanding variable lain. Namun variable ini tidak memiliki keterikatan dengan variable lain sehingga berdiri secara independent.

Akhirnya kami memutuskan untuk research mendalam terkait SOP Keselamatan Jika Terjadi Bencana. Kami menemukan bahwa *Birkenhead drill* merupakan prosedur keselamatan yang populer di tahun 1852. Pada tahun 1920 ketika terjadi peristiwa Titanic prosedur ini dilakukan oleh awak kapal, namun terjadi juga kesalahan penafsiran dari SOP tersebut [3,4] yang mengakibatkan perempuan dan anak-anak memiliki survivalbility rate yang sangat tinggi .

## Chart, line chart Description automatically generatedRandom Forest

Figure 10. Learning Curve Evalution  
  
  
Text

Description automatically generated

Figure 11. Machine Learning Output

Dari Figure 11 dan Figure 10 kita dapat melihat bahwa performa dari machine learning dengan *positive features* memiliki performa yang kurang baik dibanding dengan features yang lain. *Negative* dan *Positive features* performa yang lebih baik dibanding keduanya dan memiliki nilai yang hampir sama. Hal ini juga bisa disebabkan karena variable Sex dan Pclass berada pada Negative Features (Figure 5) dan memiliki pengaruh besar pada machine learning projek ini

## Prediction Output Visual

Figure 12. Prediction Output

Pada Figure 12 kita dapat melihat hasil output dari setiap feature berjumlah n=25.

# Kesimpulan

Dari hasil analisa yang kami lakukan terhadap dataset seluruh penumpang kapal Titanic, kami dapat mengambil beberapa kesimpulan. Dari Basic Analysis dan machine learning, kami mendapati bahwa masing-masing atribut memiliki berbagai macam faktor yang mempengaruhi tingkat keselamatan penumpang.

Semoga kedepannya dengan data yang ada kita mampu membuat system keselamatan yang lebih baik di dunia kemaritiman sehingga dapat mencegah insiden RMS Titanic terjadi lagi.

##### References

[1] A. G. Kosheleva Vladik Kreinovich, and Olga, “Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation”.

[2] M. Hardy, “Pareto’s Law,” *The Mathematical Intelligencer*, vol. 32, no. 3, pp. 38–43, Jul. 2010, doi: 10.1007/s00283-010-9159-2.

[3] M. Debczak, “The Origins of ‘Women and Children’ First,” *Mental Floss*, Apr. 19, 2022. [Online]. Available: <https://www.mentalfloss.com/posts/women-and-children-first-origins-titanic>

[4]B. S. Frey, D. A. Savage, and B. Torgler, “Noblesse oblige? Determinants of survival in a life-and-death situation,” Journal of Economic Behavior & Organization, vol. 74, no. 1–2, pp. 1–11, May 2010, doi: 10.1016/j.jebo.2010.02.005.

[5]J. Nithyashri and G. Kulanthaivel, “Classification of human age based on Neural Network using FG-NET Aging database and Wavelets,” 2012 Fourth International Conference on Advanced Computing (ICoAC), Dec. 2012, doi: 10.1109/icoac.2012.6416855.

‌