Analisa Faktor Keselamatan Dari Bencana Titanic

Dibuat Oleh:
2440006715 - Daniel Alexander
2440018394 - Divaldy Putra Eka
2401960763 - Thomas Amarta G
2401960284 - Jeremiah Dylan Julianto
2440012421 - Christopher Nathanael

Dosen Pembimbing:
D6198 – Fepri Putra Panghurian, S. Kom, M.T.I.

Even Semester 2021/2022

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

I. 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.

II. Metodologi

A. 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

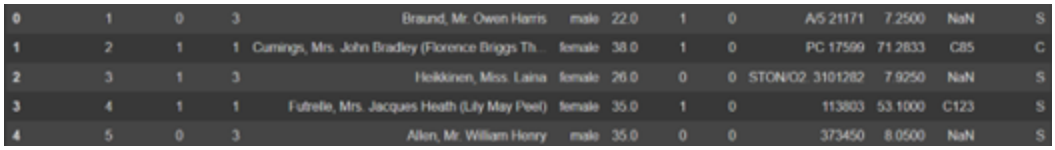
Sklearn.model_selection · **Learning_curve** · **ShuffleSplit**	· Estimate ML model performance · Split data into train and test dataset

B. 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.

Column	Data Type	Description
PassengerID	Integer	Merupakan nomor penumpang.
Survived	Integer	Menandakan apakah penumpang tersebut selamat atau tidak.
Pclass	Integer	Menandakan penumpang berada di kelas apa.
Name	Object	Merupakan nama penumpang.
Sex	Object	Merupakan gender dari penumpang.
Age	Float	Merupakan umur dari penumpang.
SibSp	Integer	Merupakan jumlah saudara.
Parch	Integer	Merupakan jumlah keluarga(orangtua/anak).
Ticket	Object	merupakan jenis tiket yang dipunyai.
Fare	Float	Merupakan harga tiket yang dibeli.
Cabin	Object	Merupakan ruang yang didapat penumpang.
Embarked	Object	Merupakan asal tempat dare penumpang.

Table 1 . Informasi Detail Dataset



0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Figure 1. Tampilan Isi Data

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

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 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.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
806	807	0	1	Andrews, Mr. Thomas Jr	male	39.0	0	0	112050	0.0	A36	S
633	634	0	1	Parr, Mr. William Henry Marsh	male	NaN	0	0	112052	0.0	NaN	S
815	816	0	1	Fry, Mr. Richard	male	NaN	0	0	112058	0.0	B102	S
263	264	0	1	Harrison, Mr. William	male	40.0	0	0	112059	0.0	B94	S
822	823	0	1	Reuchlin, Jonkheer. John George	male	38.0	0	0	19972	0.0	NaN	S
277	278	0	2	Parkes, Mr. Francis "Frank"	male	NaN	0	0	239853	0.0	NaN	S
413	414	0	2	Cunningham, Mr. Alfred Fleming	male	NaN	0	0	239853	0.0	NaN	S
466	467	0	2	Campbell, Mr. William	male	NaN	0	0	239853	0.0	NaN	S
481	482	0	2	Frost, Mr. Anthony Wood "Archie"	male	NaN	0	0	239854	0.0	NaN	S
732	733	0	2	Knight, Mr. Robert J	male	NaN	0	0	239855	0.0	NaN	S
674	675	0	2	Watson, Mr. Ennis Hastings	male	NaN	0	0	239856	0.0	NaN	S
179	180	0	3	Leonard, Mr. Lionel	male	36.0	0	0	LINE	0.0	NaN	S
271	272	1	3	Tornquist, Mr. William Henry	male	25.0	0	0	LINE	0.0	NaN	S
302	303	0	3	Johnson, Mr. William Cahoone Jr	male	19.0	0	0	LINE	0.0	NaN	S
597	598	0	3	Johnson, Mr. Alfred	male	49.0	0	0	LINE	0.0	NaN	S

Figure 3. Tampilan Ticket dan Fare

C. Data Cleansing

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

```
============
Null data
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

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

D. 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

E. 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*

F. Features

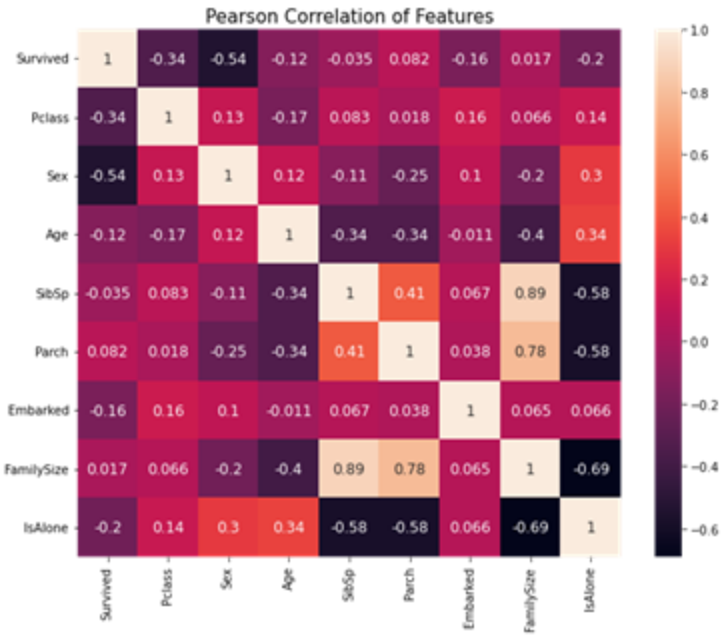


Figure 5. Heatmap

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

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

G. 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)

```
# Kita akan membagi data dengan metode 80/20 (Pareto Principle)
xtrain, xtest, ytrain, ytest = train_test_split(features,targets,test_size=0.2)
print("xinput Shape: ", xtrain.shape)
print("ytrain Shape: ", xtest.shape)
print("xtest Shape: ", ytrain.shape)
print("ytest Shape: ", ytest.shape)
```

All Features

```
# Kita akan membagi data dengan metode 80/20 (Pareto Principle)
xtrain1, xtest1, ytrain1, ytest1 = train_test_split(features,targets,test_size=0.2)
print("xinput Shape: ", xtrain1.shape)
print("ytrain Shape: ", xtest1.shape)
print("xtest Shape: ", ytrain1.shape)
print("ytest Shape: ", ytest1.shape)
```

Positive Features

```
# Kita akan membagi data dengan metode 80/20 (Pareto Principle)
xtrain2, xtest2, ytrain2, ytest2 = train_test_split(features,targets,test_size=0.2)
print("xinput Shape: ", xtrain2.shape)
print("ytrain Shape: ", xtest2.shape)
print("xtest Shape: ", ytrain2.shape)
print("ytest Shape: ", ytest2.shape)
```

Negative Features

H. 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

```
Train Score RandomForestClassifier : 0.8637640449438202
Test Score RandomForestClassifier : 0.8547486033519553
```

Positive Features

```
Train Score Positive RandomForestClassifier : 0.6783707865168539
Test Score Positive RandomForestClassifier : 0.6424581005586593
```

Negative Features

```
Train Score Negative RandomForestClassifier : 0.8595505617977528
Test Score Negative RandomForestClassifier : 0.8044692737430168
```

III. Hasil

A. 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.

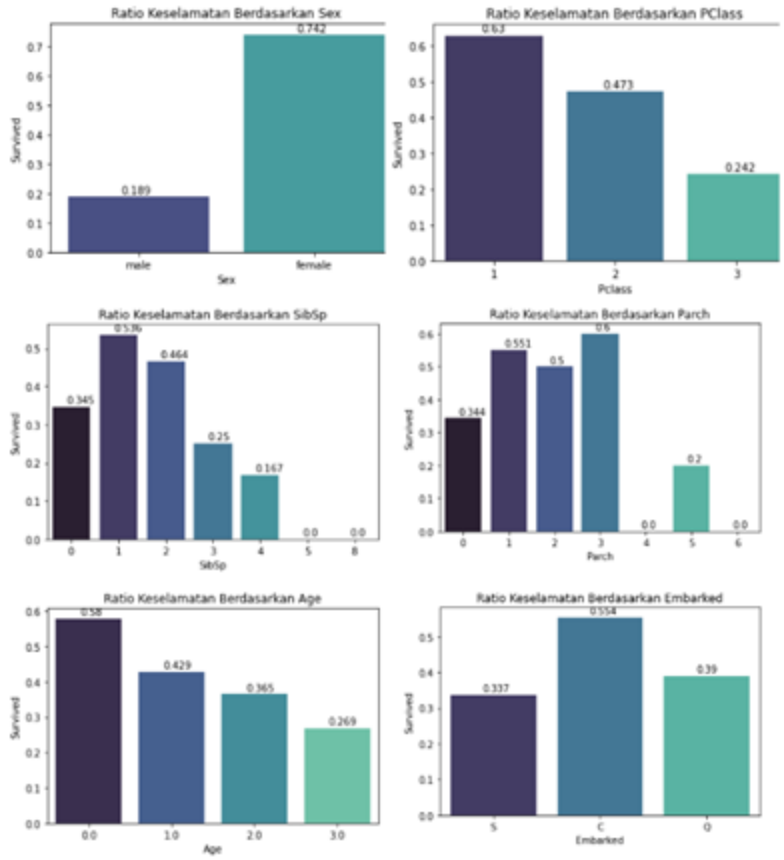


Figure 6. Hasil Basic Analysis

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.

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

i) SibSp dan Parch

Family size dan IsAlone

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

```
#Family Size
dfTrain['FamilySize'] = dfTrain['SibSp']+dfTrain['Parch']+1
# +1 adlaah subject / passanger  yang bersangkutan
```

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

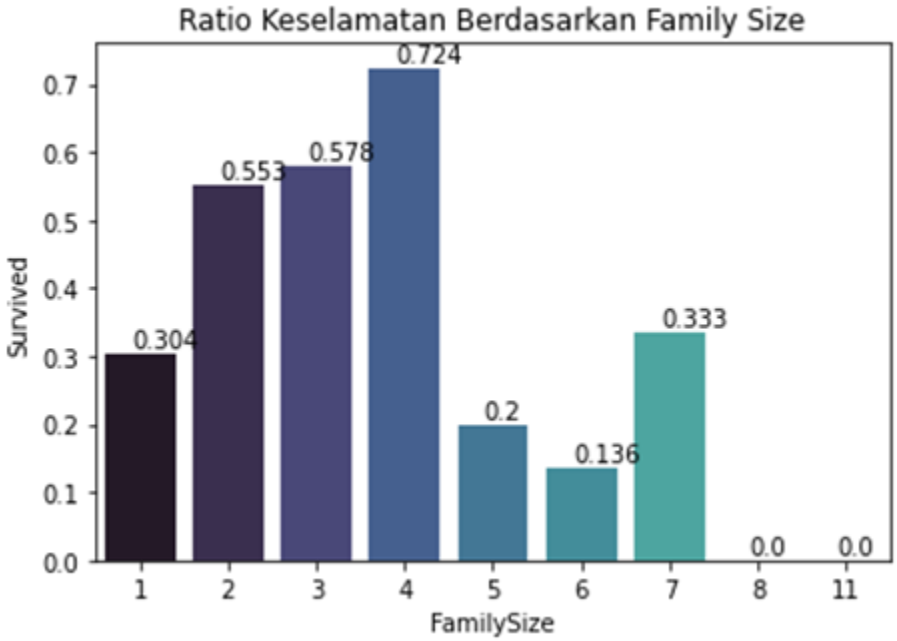


Figure 7. Ratio Family Size Terhadap Survied

Pada 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.

```
#IsAlone
dfTrain['IsAlone']=1
dfTrain['IsAlone'].loc[dfTrain['FamilySize']>1]=0
# IsAlone =1 jika Passanger memiliki nilai SibSp dan Parch 0
```

Figure 8. Ratio Size IsAlone Terhadap Survived

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

ii) Embarked

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

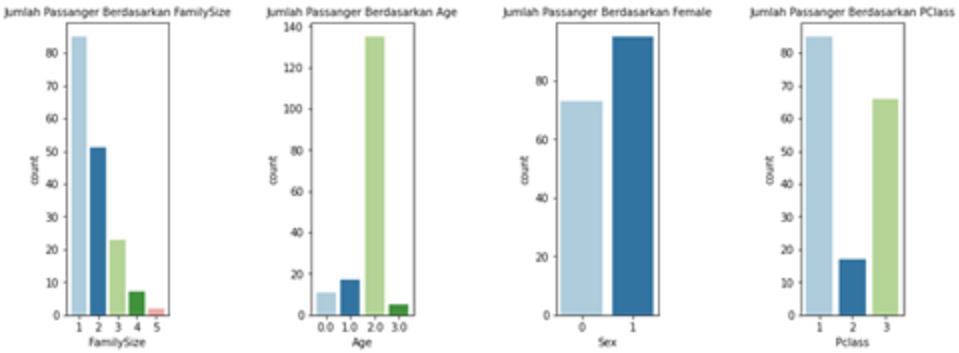


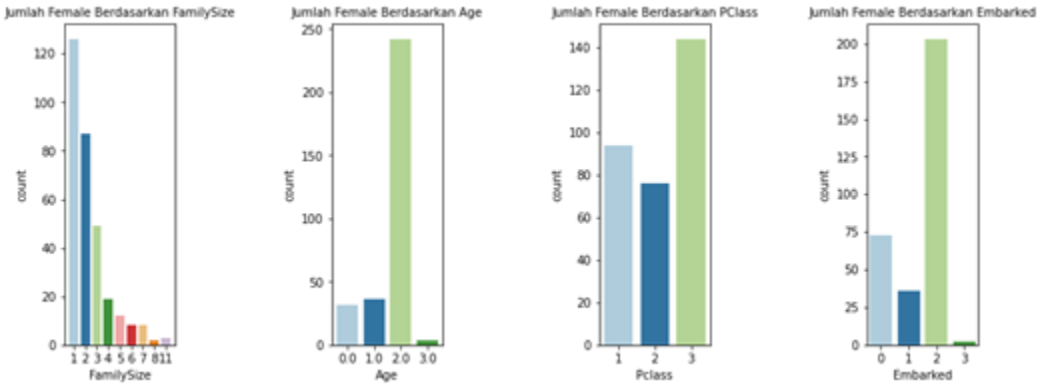
Figure 9. Countplot Embarked (C)

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

iii) Sex



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 .

C. Random Forest

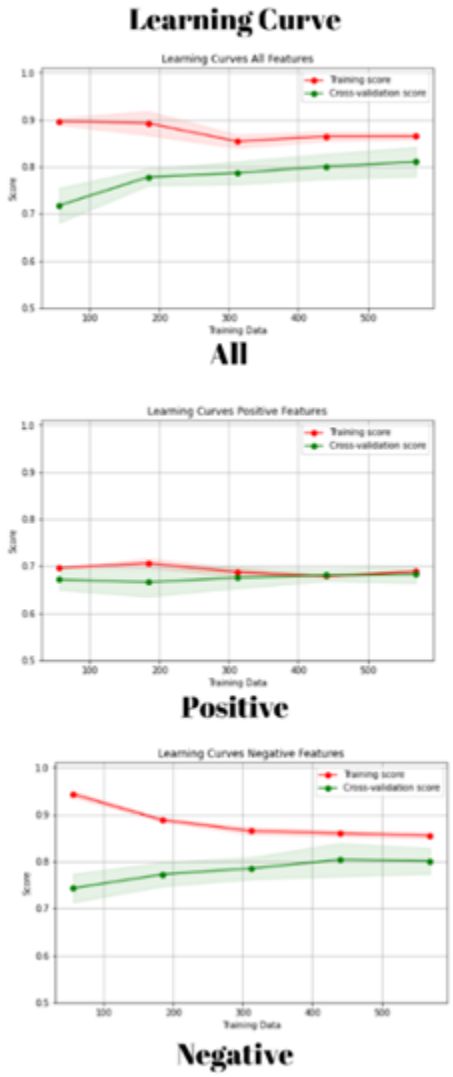
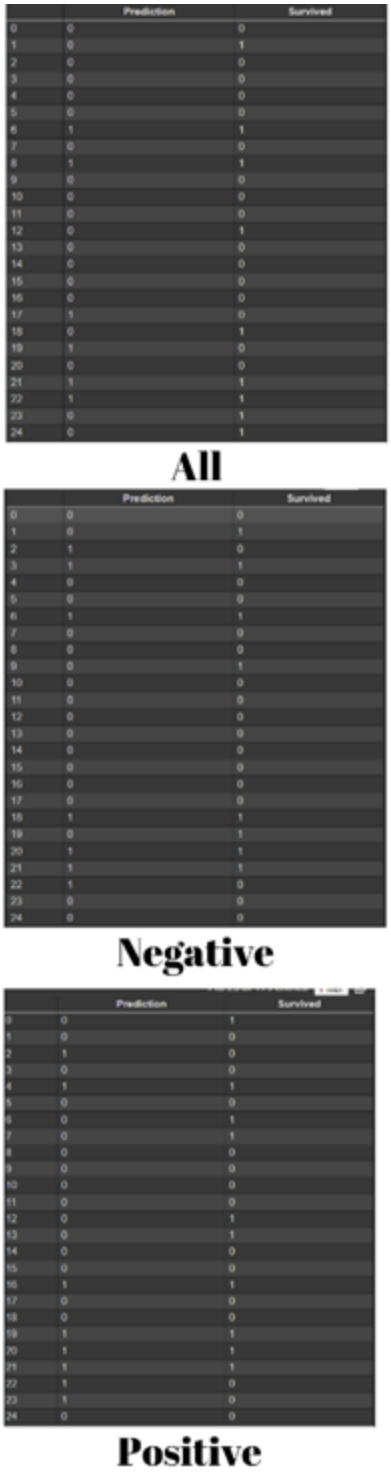


Figure 10. Learning Curve Evalution

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

D. Prediction Output Visual



	Prediction	Survived
0	0	0
1	0	1
2	0	0
3	0	0
4	0	0
5	0	0
6	1	1
7	0	0
8	1	1
9	0	0
10	0	0
11	0	0
12	0	1
13	0	0
14	0	0
15	0	0
16	0	0
17	1	0
18	0	1
19	1	0
20	0	0
21	1	1
22	1	1
23	0	1
24	0	1

	Prediction	Survived
0	0	0
1	0	1
2	1	0
3	1	1
4	0	0
5	0	0
6	1	1
7	0	0
8	0	0
9	0	1
10	0	0
11	0	0
12	0	0
13	0	0
14	0	0
15	0	0
16	0	0
17	0	0
18	1	1
19	0	1
20	1	1
21	1	1
22	1	0
23	0	0
24	0	0

	Prediction	Survived
0	0	1
1	0	0
2	1	0
3	0	0
4	1	1
5	0	0
6	0	1
7	0	1
8	0	0
9	0	0
10	0	0
11	0	0
12	0	1
13	0	1
14	0	0
15	0	0
16	1	1
17	0	0
18	0	0
19	1	1
20	1	1
21	1	1
22	1	0
23	1	0
24	0	0

Figure 12. Prediction Output

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

IV. 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.