*Laporan UTS - IBDA3111*

Calvin Institute of Technology

Semester ganjil 2022/2023



Oleh

Stefannus Christian / 202000138 / IT & Big Data Analytics

Judul: Pembersihan Dataset Titanic dan membandingkan beberapa Model Klasifikasi

Disclaimer: Pada proposal data, saya menyebutkan bahwa saya akan membuat model Logistic Regression untuk memprediksi apakah suatu penumpang titanic selamat atau tidak. Pada laporan kali ini, saya akan membandingkan dan memaparkan 8 model lain selain Logistic Regression.

Latar Belakang

Tenggelamnya Titanic adalah salah satu *shipwreck* paling terkenal dalam sejarah. Pada tanggal 15 April 1912, selama pelayaran perdananya, RMS Titanic yang secara luas dianggap "tidak dapat tenggelam" tenggelam setelah bertabrakan dengan *iceberg*. Sayangnya, tidak ada cukup sekoci untuk semua penumpang, mengakibatkan kematian 1502 dari 2224 penumpang dan awak. Meskipun ada beberapa unsur keberuntungan yang terlibat dalam bertahan hidup, tampaknya beberapa kelompok orang lebih mungkin untuk bertahan hidup daripada yang lain.

Deskripsi Data

Terdapat 12 fitur dari dataset titanic yang saya ambil. Terdapat 11 fitur input dan 1 fitur

output.

1. PassengerID: Passenger ke i dari dataset (Unique Value).

2. Survived: Menandakan apakah suatu penumpang selamat atau tidak. 0 = tidak

selamat, 1 = selamat (Merupakan fitur output).

3. Pclass (Passenger Class): Divisi dari penumpang. Terdiri dari 3 kelas yaitu kelas 1, 2,

dan 3. Deskripsi Kelas: 1 → Upper Class, 2 → Middle Class, 3 → Lower Class.

4. Name: Nama dari penumpang.

5. Sex: Jenis kelamin penumpang.

6. Age: Umur dari penumpang.

7. SibSp (Sibling Spouse): Jumlah saudara kandung atau pasangan dari seorang

penumpang di dalam kapal. Note: Mistresses and Fiances Ignored.

8. Parch: Jumlah orangtua atau anak dari seorang penumpang di dalam kapal.

9. Ticket: Nomor ticket dari penumpang.

10. Fare: Harga tiket kapal yang dibayar penumpang dalam British Pound pada saat itu.

11. Cabin: Nomor Cabin penumpang.

12. Embarked: Dari mana penumpang naik kapal. Ada tiga kemungkinan nilai untuk fitur

ini yaitu Southampton, Cherbourg, dan Quennstown

Sumber Data

Data dapat diambil dari sumber berikutL <https://github.com/krishnaik06/EDA1/blob/master/titanic_train.csv>

Dataset

Graphical user interface, text

Description automatically generated

Dapat dilihat dari dataset bahwa terdapat data – data kosong dan banyak data kategorikal sehingga diperlukan prapemrosesan dan rekayasa data sebelum dimasukkan ke *Machine Learning Model.*

Prapemrosesan / Rekayasa Data

Mendrop Kolom PassengerID dan Ticket & Mengganti Nama Kolom Dataset

Text

Description automatically generated with low confidence

Kolom PassengerId isinya hanyalah id penumpang dengan kata lain menandakan penumpang ke – i sehingga tidak relevan bagi model dan dapat dibuang. Kemudian, saya juga mendrop kolom ticket karena isi dari kolom ticket adalah nomor ticket dari penumpang. Sulit untuk mendapatkan sesuatu yang berguna dari kolom tersebut, sehingga saya memutuskan untuk membuang kolom tersebut.

Menambahkan Fitur Family Size

Graphical user interface, text, website

Description automatically generated

Setelah melihat dari beberapa sumber, saya melihat ada satu sumber yang menambahkan fitur family size yaitu jumlah spouse ditambah dengan jumlah orangtua/anak yang terdapat pada kapal, sehingga saya memutuskan untuk menambahkan fitur ini ke dataset.

Menambahkan Title dari nama masing – masing penumpang

A screenshot of a computer

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generated

Saya membuat fungsi menggunakan *regex (regular expression)* untuk mendeteksi *title* penumpang seperti Mr, Miss, Dr, dll. Tujuan saya melakukan hal ini adalah agar fitur nama tidak dibuang begitu saja. Saya melihat adanya titel – titel ini yang dapat diekstrak menjadi fitur baru yang mungkin berguna ketika pelatihan model nanti.

Dataset setelah menambahkan fitur Title

Graphical user interface, application

Description automatically generated

Membuang Kolom Nama

Text

Description automatically generated

Informasi penting yang dapat diambil dari nama hanyalah titel dari penumpang sehingga kolom nama dapat di buang.

Mengelompokkan Fitur *Fare* menjadi 4 kelompok

A screenshot of a computer

Description automatically generated with medium confidence

Saya mengelompokkan fitur *Fare* menjadi 4 kelompok karena fitur *fare* memiliki range data yang berbeda dari data – data lainnya. Data – data lainnya akan berbentuk kategorikal (0 – 1) dan fitur *fare* tidak berbentuk kategorikal, sehingga saya rasa bahwa fitur *fare* akan merusak pemodelan. Maka dari itu, memutuskan untuk mengelompokkan fitur *Fare* agar nantinya fitur ini dapat berbentuk kategorikal juga.

Dataset setelah merubah Fitur *Fare* menjadi 4 kelompok

Graphical user interface, text

Description automatically generated

Metode Data Cleaning 1 – Membuang Data Duplikasi

A picture containing table

Description automatically generated



Graphical user interface, text

Description automatically generated

Terdapat 194 baris data duplikasi yang saya buang. Saya membuang duplikasi data sehingga model dapat digeneralisasi dengan lebih baik ke *dataset*.

Metode Data Cleaning 2 – Imputasi Data

Imputasi data dilakukan karena ada 4 fitur dataset yang memiliki data – data null.

Graphical user interface

Description automatically generated

Drop Fitur Cabin

**Text

Description automatically generated**

**Dari 697 baris data tersisa, 494 (≈71%) baris dari fitur *Cabin* merupakan data kosong, dan sulit untuk melakukan imputasi pada fitur ini sehingga saya memutuskan untuk membuang fitur *Cabin*.**

Imputasi Fitur Age berdasarkan kelas Penumpang

Chart, box and whisker chart

Description automatically generated

Untuk imputasi fitur *Age*, saya memutuskan untuk menggunakan imputasi rata – rata. Tetapi, rata – rata yang saya gunakan bukan satu rata – rata saja tetapi tiga rata – rata. Tiga rata – rata ini saya klasifikasikan berdasarkan class penumpang (penjelasan di gambar bawah).

Text

Description automatically generated

Berikut adalah *source code* yang saya buat untuk mengimputasi data – data null pada fitur *age* berdasarkan *Passenger Class*.

Text

Description automatically generated

Graphical user interface, text

Description automatically generated with medium confidence

Setelah fitur *Age* diimputasi, dapat dilihat bahwa sudah tidak ada lagi baris yang null pada fitur *Age*.

Mengelompokkan Fitur Age menjadi 4 Kelompok

Graphical user interface, text

Description automatically generated

Saya mengelompokkan Fitur *Age* karena sama dengan *Fare* takutnya karena fitur *Age* tidak bersifat kategorikal (0-1, dll) maka saya takutnya fitur ini akan “merusak” pelatihan model sehingga saya mengelompokkan fitur ini sehingga nantinya dapat dilakukan *One Hot Encoding*.

A picture containing text

Description automatically generated

Imputasi Kolom Embarked

Text

Description automatically generated

Hanya terdapat 2 baris dari fitur *Embarked* yang memiliki data null sehingga saya akan mengimputasi kolom embarked ini menggunakan modus atau data terbanyak. Data terbanyak fitur *Embarked* adalah *Southampton* sehingga saya akan mengisi 2 baris yang kosong ini dengan *Southampton.*

Mengatasi Fitur Fare yang memiliki null *values*

Saya akan mengatasi null *values* pada fitur *fare* dengan One Hot Encoding. One Hot Encoding akan mengimputasi sendiri data – data yang null pada fitur *fare.*

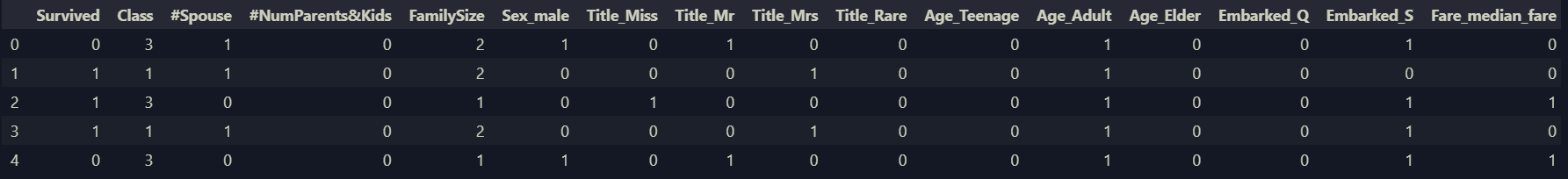
One Hot Encoding

Text

Description automatically generated

Saya melakukan *One Hot Encoding* untuk fitur – fitur *Sex, Title, Age, Embarked, dan Fare*.

Berikut adalah tampilan dataset setelah dilakukan One Hot Encoding.



Summary DataFrame setelah di One Hot Encode

Graphical user interface

Description automatically generated with medium confidence

Text

Description automatically generated

Gambar diatas adalah untuk pengecekan tipe data dari dataframe setelah dilakukan one hot encoding. Diperlukan semua data – data numerik agar model dapat dilatih sehingga perlu dipastikan terlebih dahulu kalau semua fitur dataframe adalah data – data numerik.

Correlation Matrix

Berikut adalah matriks korelasi dataframe setelah dilakukan one hot encoding.

Chart, treemap chart

Description automatically generated

RFE Feature Selection

Text

Description automatically generated

Membandingkan Performa 9 Model Klasifikasi sebelum dan sesudah dilakukan RFE Feature Selection

A picture containing text

Description automatically generated

Detail kode dapat dilihat pada file ipynb. RFE estimator terbaik yang saya dapatkan adalah Gradient Boosting estimator dengan model Logistic Regression.

Graphical user interface, text

Description automatically generated

Ini harusnya ada pada tabel peningkatan akurasi. 

Dapat dilihat dari kode diatas bahwa dari ke sembilan model, terdapat 8 model yang memiliki peningkatan akurasi ketika dilakukan RFE Feature estimator dengan menggunakan estimator Gradient Boosting dan mengambil 11 fitur.

Dapat dilihat bahwa model KNN dengan 4 neighbors memiliki akurasi terbaik setelah dilakukan RFE Feature Selection yaitu sebesar 81.429%.



Mutual Information Feature Selection

Detail kode dapat dilihat pada file ipynb

Text

Description automatically generated

Saya menggunakan model KNN dengan n neighbors = 4 berdasarkan hasil dari RFE feature selection dan dapat dilihat bahwa metode mutual information mengambil 7 fitur terbaik dengan akurasi 80%.

Solusi

* Gunakan model KNN untuk dataset Titanic.
* Gunakan Feature Selection RFE dengan estimator Gradient Boosting dan Model Logistic Regression dan lakukan for loop untuk membandingkan semua n\_features\_to\_select dan gunakan lah hasil akurasi yang terbaik
* Lakukan lah one hot encoding dan pecahlah kolom age dan fare seperti yang saya lakukan diatas
* Jangan langsung Drop Fitur Nama tetapi ambilah title dari nama tersebut
* Buanglah Kolom Cabin, PassengerId, dan Ticket

Kesimpulan

* Model KNN dengan n\_neighbors = 4 merupakan model terbaik untuk mengklasifikasi apakah suatu penumpang selamat atau tidak dari dataset titanic.
* Akurasi Model KNN setelah dilakukan RFE Feature Selection dengan memilih 11 fitur dan estimator gradient boosting adalah 81%.
* Feature selection meningkatkan akurasi model untuk titanic Dataset
* RFE Feature Selection lebih baik daripada mutual Information Feature Selection

Komitmen Integritas

“Di hadapan TUHAN yang hidup, saya menegaskan bahwa saya tidak memberikan

maupun menerima bantuan apapun—baik lisan, tulisan, maupun elektronik—di dalam

ujian ini selain daripada apa yang telah diizinkan oleh pengajar, dan tidak akan

menyebarkan baik soal maupun jawaban ujian kepada pihak lain.”



Stefannus Christian 202000138