Mata Kuliah - Penggalian Data

**Nama Kelompok** : [dikosongi]

**Anggota** :

[201910370311337 - SUKMA]

[202110370311135 - AKMAL SHAHIB MAULANA]

Berikut ini merupakan update template laporan Mini Project kuliah Penggalian Data.

**Nilai Total: 120 poin**

**Tahap 0 (poin: 25)**: Business Objective

Business objective pada projek ini adalah untuk mencari informasi mengenai faktor apa saja yang bisa menjadikan suatu film menjadi populer.

**Tahap 1 (poin: 25)**: Original Data

* Topik ini kami ambil karena melihat kondisi industri film dan series Indonesia yang masih dalam kondisi berkembang. Kebanyakan dari film Indonesia hanya dapat menjadi laris dan banyak diminati karena sang sutradara film nya seperti misalnya atau karena pemerannya. Dari beberapa kasus tidak sedikit yang seketika langsung tidak tertarik pada suatu film karena poster nya atau trailer nya padahal filmnya bagus atau sebaliknya.
* Data yang digunakan.
  + Dataset kami berisi informasi lengkap tentang berbagai judul yang tersedia di Netflix, termasuk judul, tahun rilis, runtime, skor IMDB, dan suara IMDB
  + Sebutkan dan jelaskan atribut pada data tersebut.
    - judul: Nama acara TV atau film. (Teks)
    - type: Menunjukkan apakah entri adalah acara TV atau film. (Teks)
    - deskripsi: Ringkasan atau deskripsi singkat tentang acara TV atau film. (Teks)
    - tahun\_rilis: Tahun rilis acara TV atau film. (Numerik)
    - age\_certification: Peringkat sertifikasi usia untuk acara TV atau film. (Teks)
    - runtime: Durasi episode untuk acara TV atau durasi film. (Teks)
    - imdb\_id: id dari imdb
    - imdb\_score: Skor yang diberikan oleh pengguna di IMDB untuk judul tertentu. (Numerik)
    - imdb\_votes: Jumlah suara yang diterima oleh setiap judul di IMDB. (Numerik)
  + Jelaskan data mining task yang akan digunakan (*classification, clustering, regression, association rule mining, anomaly detection,* dsb.).
    - regression : kami juga akan menetapkan variabel-variabel independen dan variabel dependen kemudian akan kami teliti variabel manakah mempunyai pengaruh besar.

tak menutup kemungkinan kami akan melakukan data mining tasks lainnya jika seandainya diperlukan untuk mendapatkan hasil yang lebih baik.

* Sumber data untuk proyek ini kami dapatkan dari Kaggle yang bisa diakses melalui [tautan ini](https://www.kaggle.com/datasets/thedevastator/netflix-imdb-scores).

**Tahap 2 (poin: 10)**: Target Data (Optional)

* Poin berikut merupakan atribut dari data yang akan kami gunakan:
  + type
  + release\_year
  + age\_certification
  + runtime
  + imdb\_score

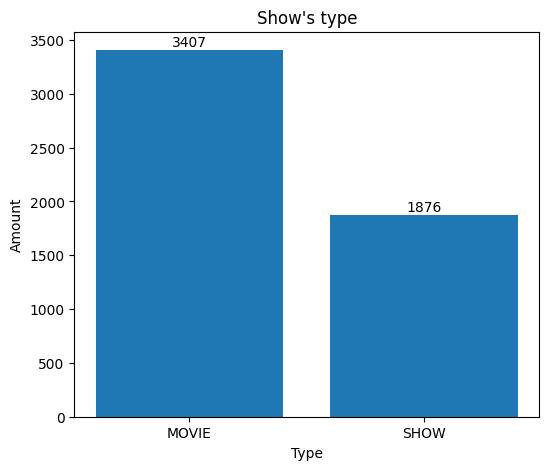
Kami juga menyertakan sedikit preview dari data kami sebagai berikut :

Jumlah film berdasarkan tahunnya :

da

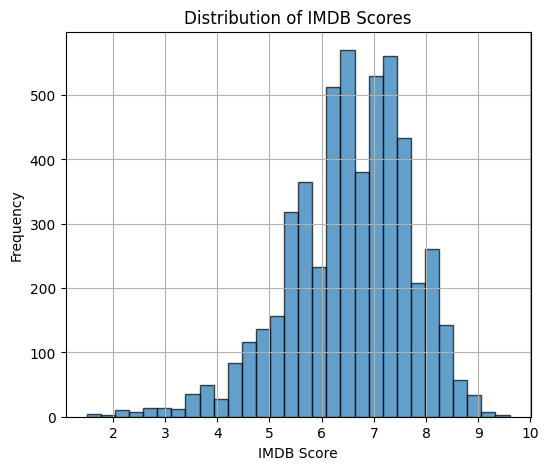

gambar 1 : jumlah film dari tahun ketahun

Jumlah film berdasarkan tipe :



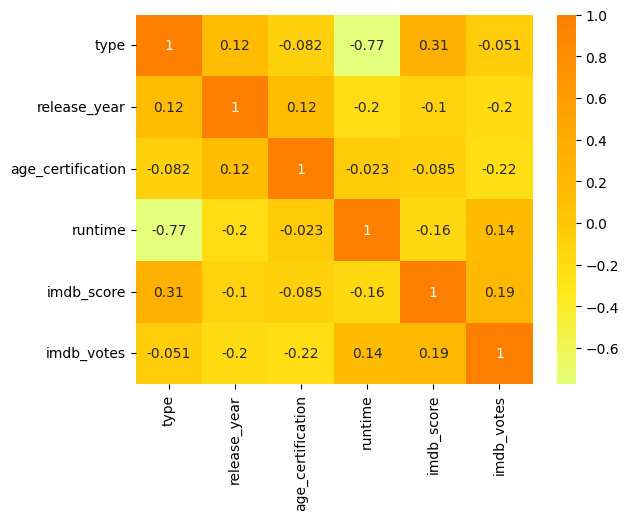
gambar 2 : jumlah tontonan bertipe film dan TV show

Distribusi data pada kolom imdb score



gambar 3 : distribusi data rating dari film dan TV show

Tabel korelasi :



gambar 4 : Tabel korelasi

Pada gambar 1 diperlihatkan bahwa jumlah film paling banyak meningkat pada 2010 hingga tahun 2020. Gambar 2 menunjukan bahwa ada ketidak seimbangan pada jumlah data antara tipe film dan TV show. Hal yang sama ditampilkan pada gambar 3 dimana banyak data pada rentang skor 8 sedangkan dengan nilai 5 kebawah tergolong sedikit.

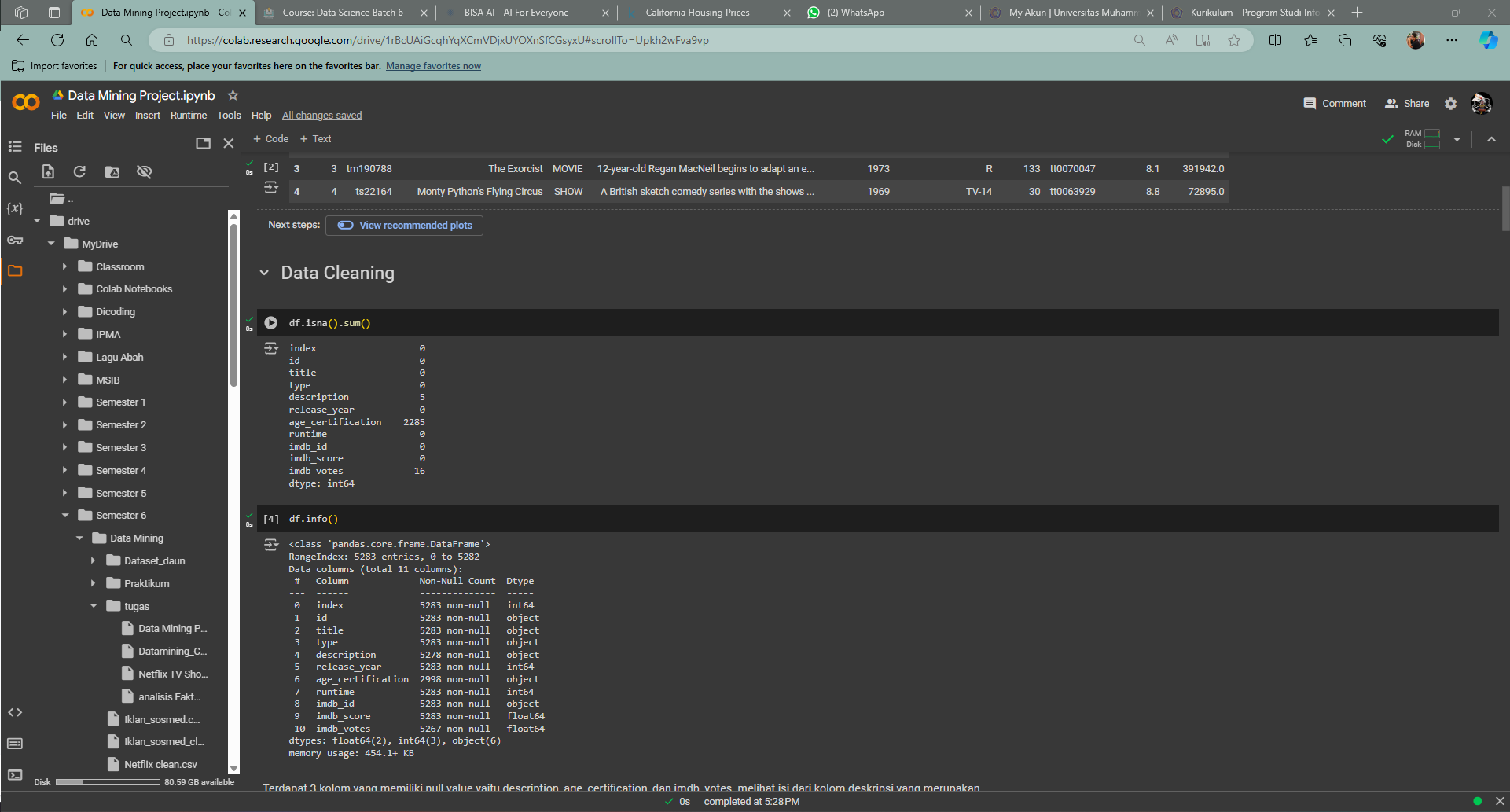
**Tahap 3-4 (poin: 25)**: Data Pre-processing & Transformation

Beberapa teknik yang bisa digunakan yaitu (tentu sesuai kondisi dan kebutuhan):

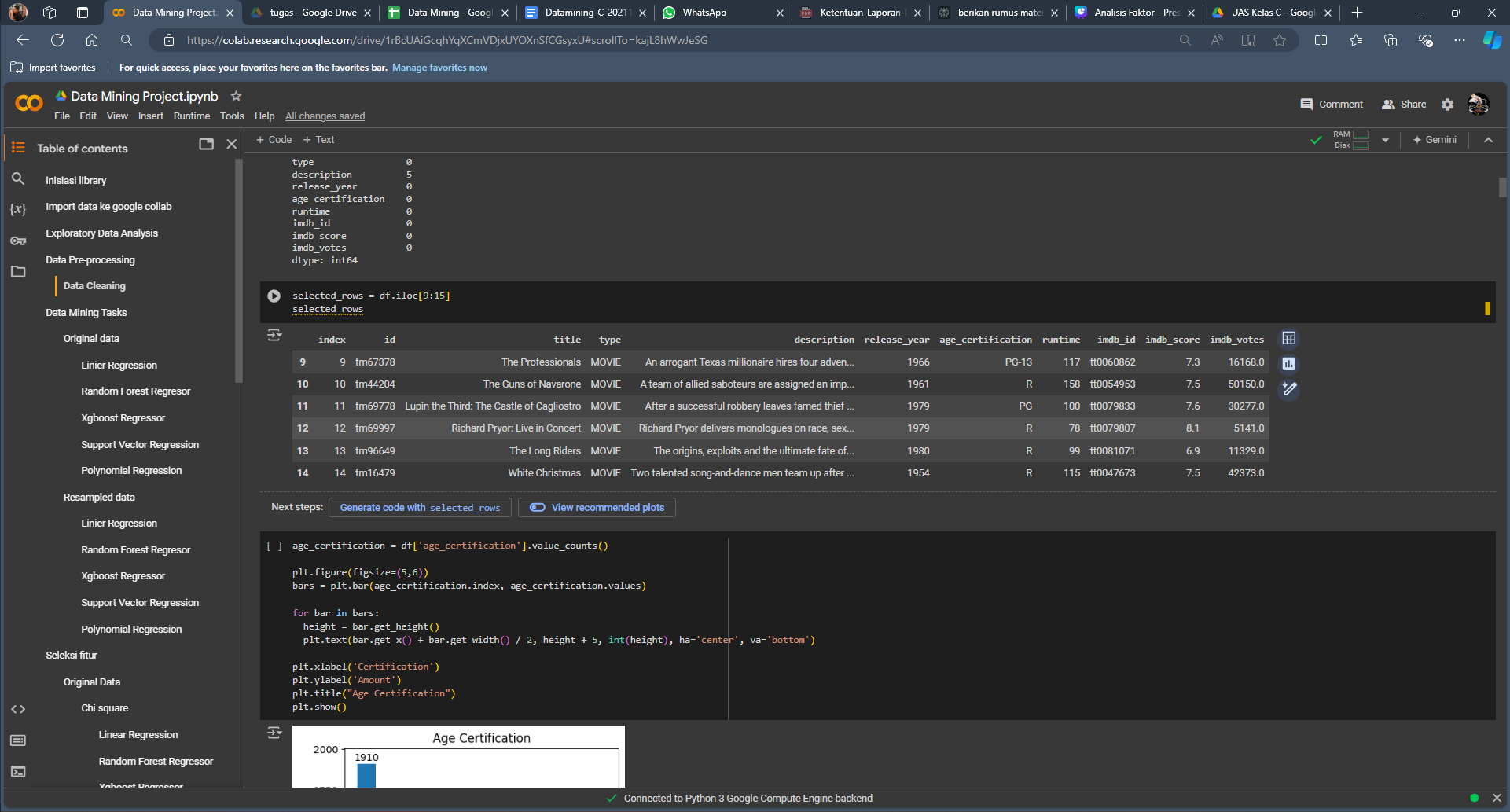
* Data Cleaning (and/or data correcting)

Before :

Terdapat 3 kolom yang memiliki null value yaitu description, age\_certification, dan imdb\_votes. melihat isi dari kolom deskripsi yang merupakan data teks dan masuk ke ranah NLP maka null value dari kolom ini tidak akan kami gunakan pada project ini. Sementara untuk 2 kolom lain nya yang merupakan data kategorikal yang bisa di encode menjadi angka maka akan kami berlakukan data cleaning

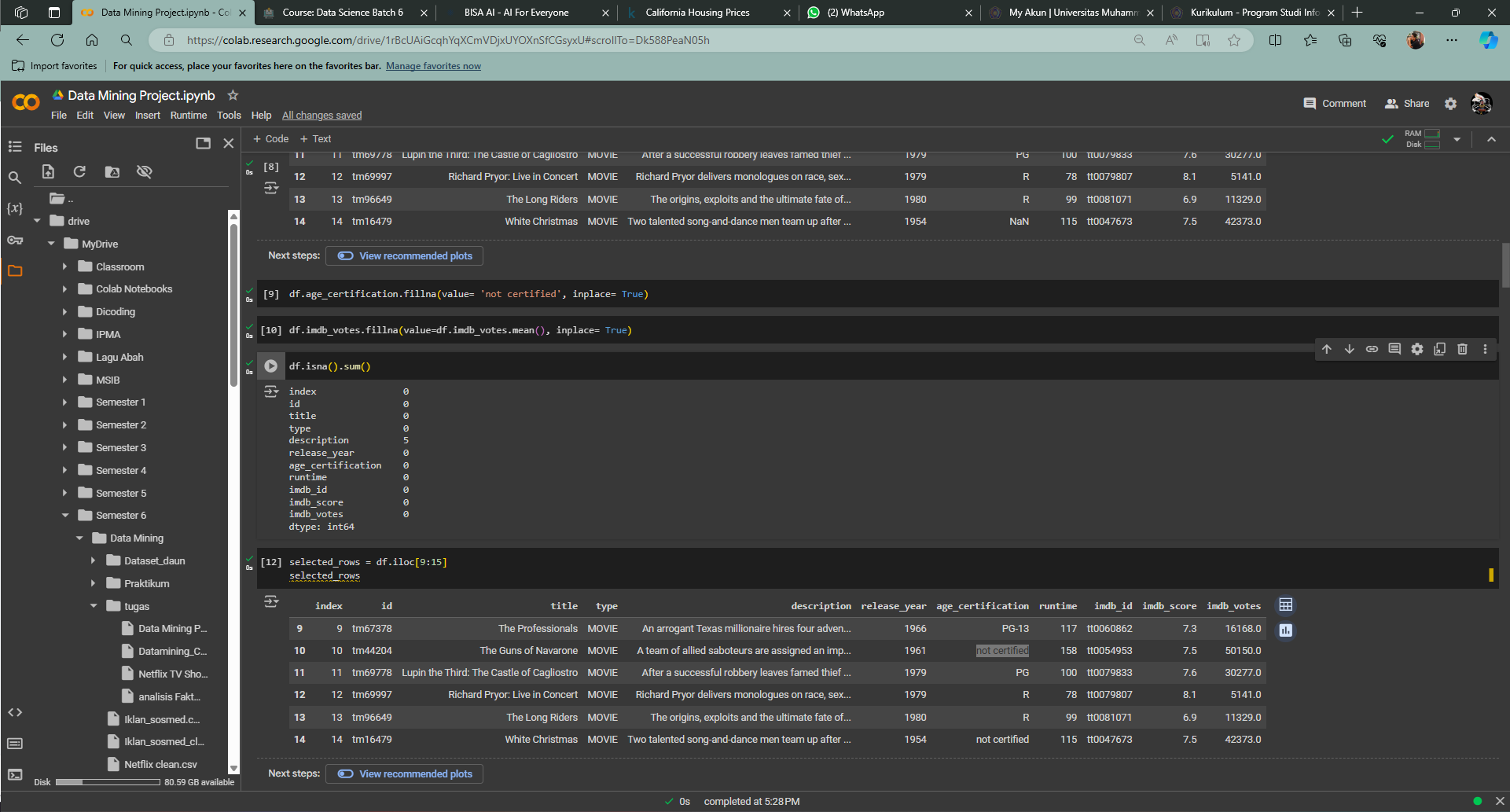


gambar 5 : daftar fitur yang ada pada dataset dengan jumlah kolom kosong

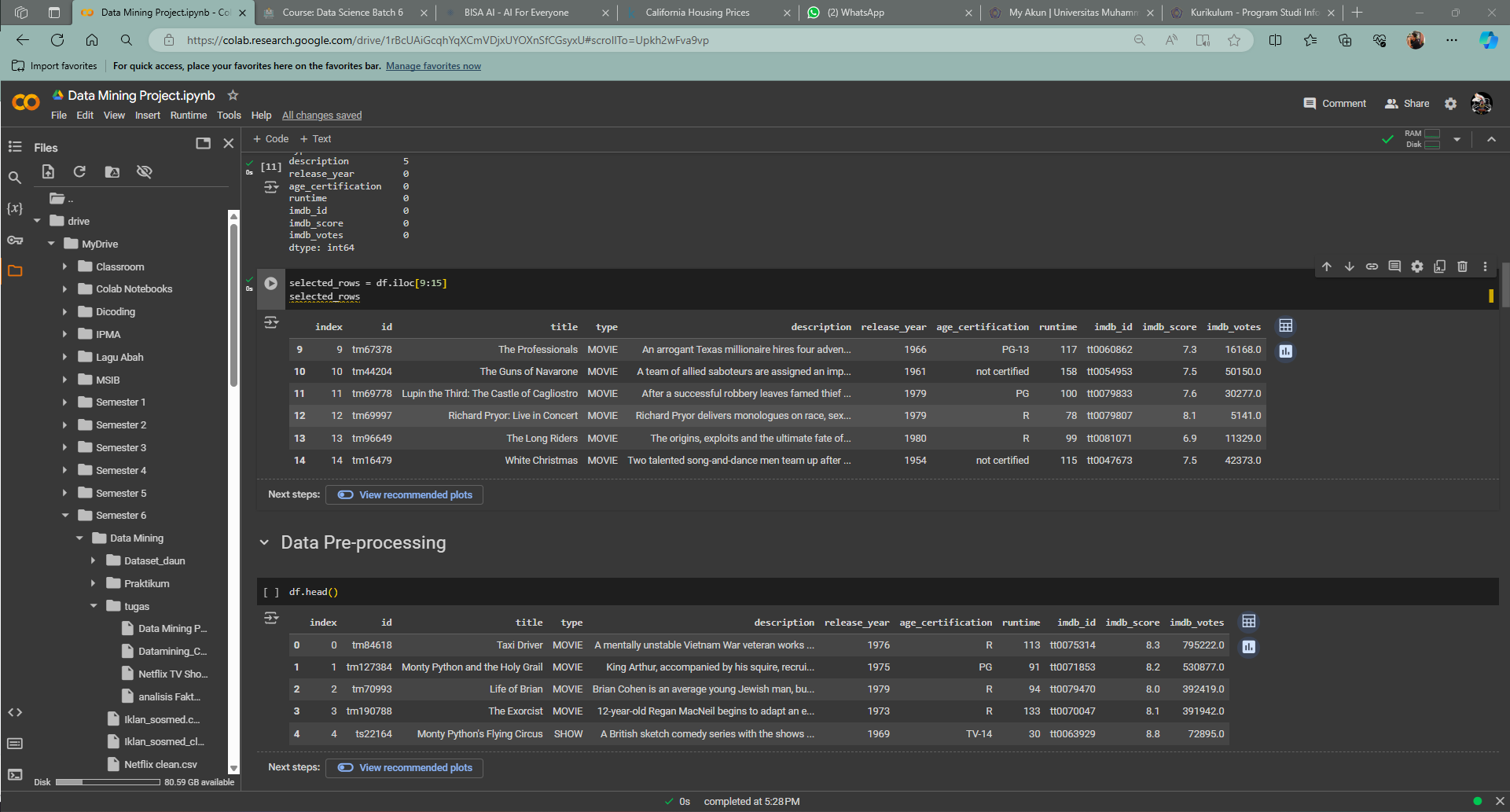


gambar 6 : preview data dengan kolom kosong

After data cleaning :



gambar 7 : daftar fitur dan kolom setelah dilakukan data cleaning



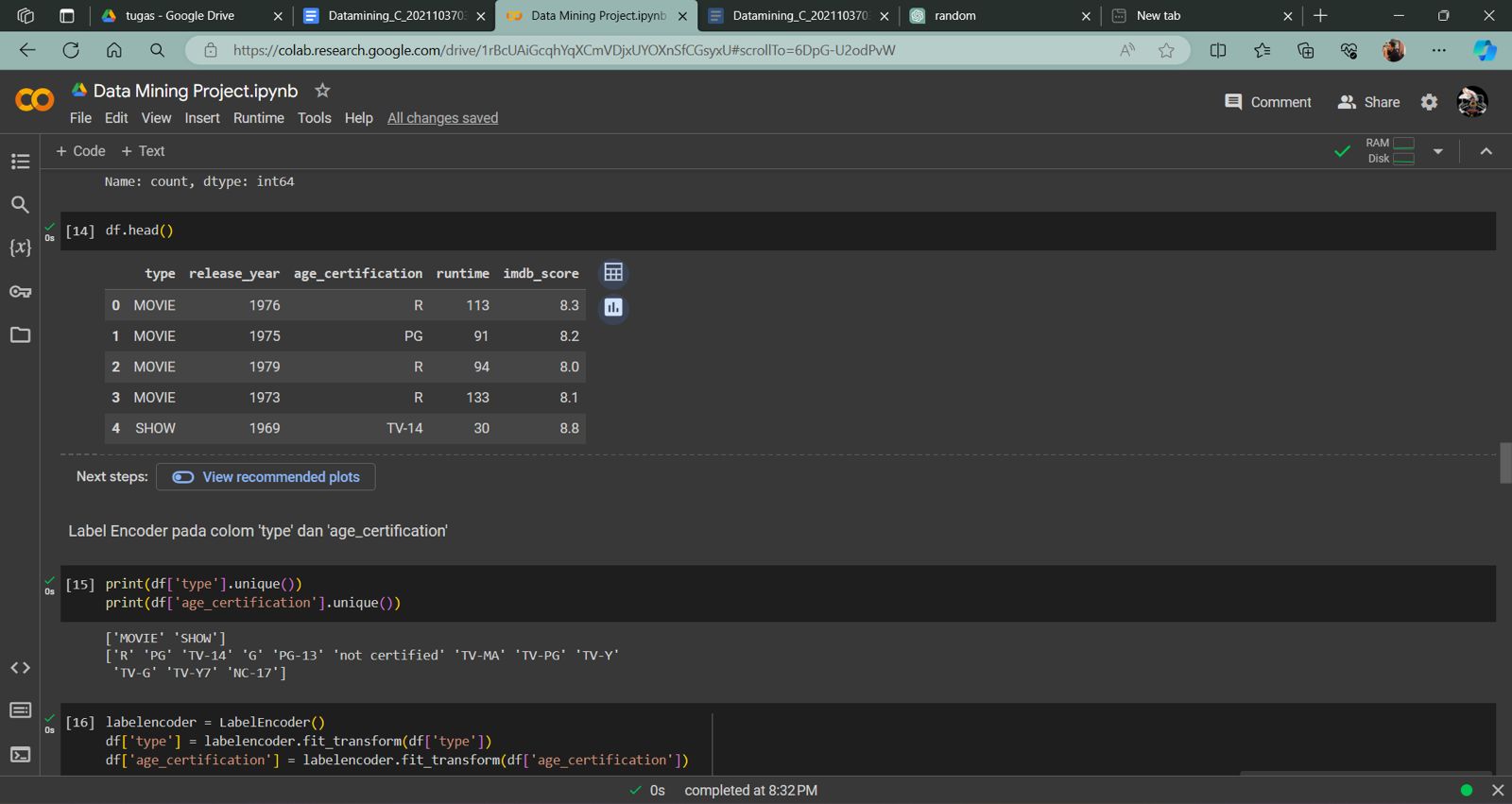
gambar 8 : preview data setelah data cleaning

Bisa dilihat gambar 5 ada daftar atribut pada dataset beserta angka disampingnya. Angka ini menunjukan ada berapa baris yang kosong. Ada 3 atribut yang disampingnya adalah angka selain 0. Kemudian pada gambar 7 angkanya berubah menjadi 0. Pada indeks ke-10 dan ke-14 yang awalnya berisi “Nan” yang menandakan tidak ada nilai berubah menjadi nilai random

* Data Transformation

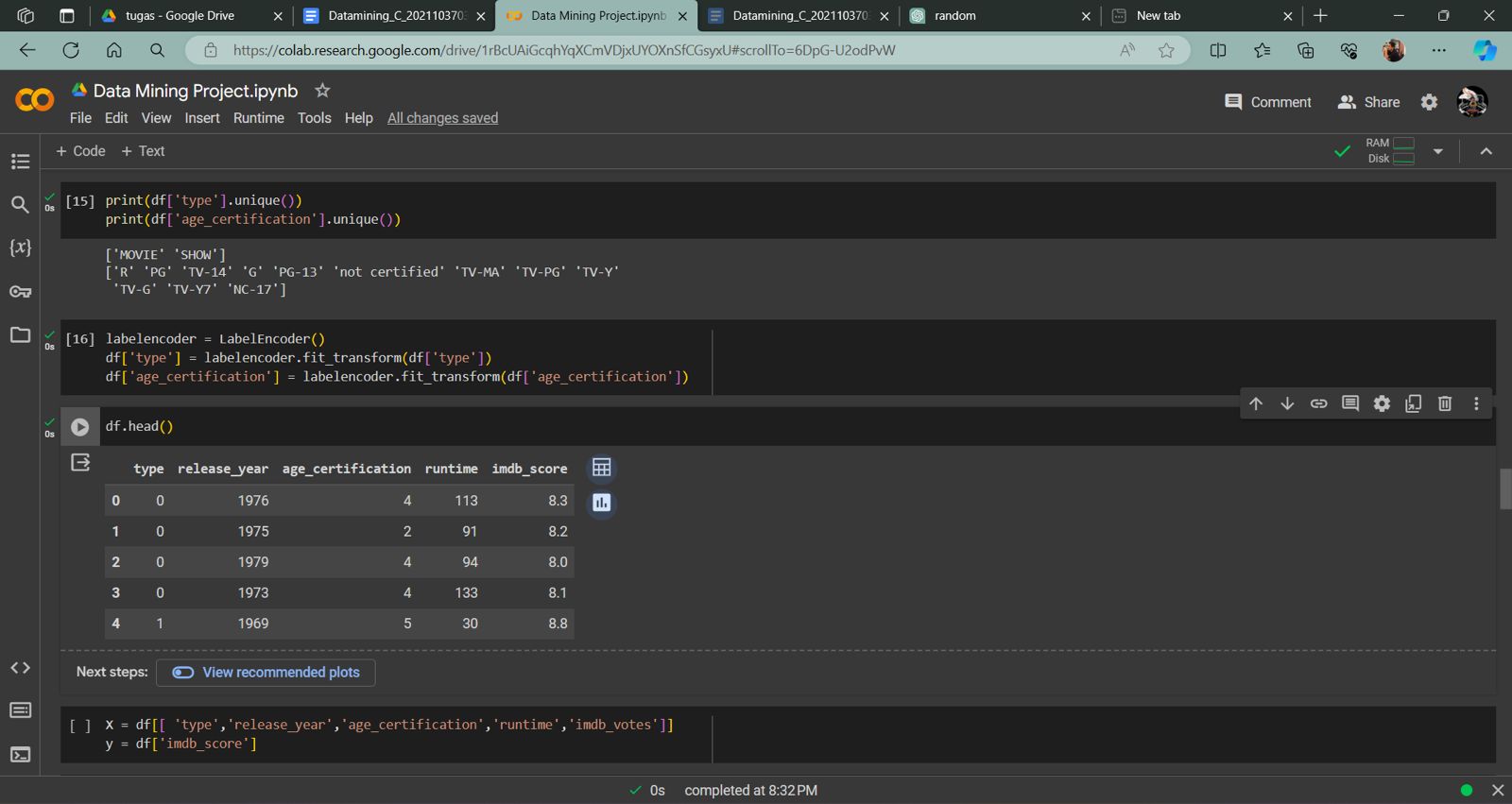
kami merubah value dari beberapa kolom yang memiliki tipe data string atau object menggunakan label encoder

Before



gambar 9 : Data target yang masih dalam bentuk kategorikal

After :

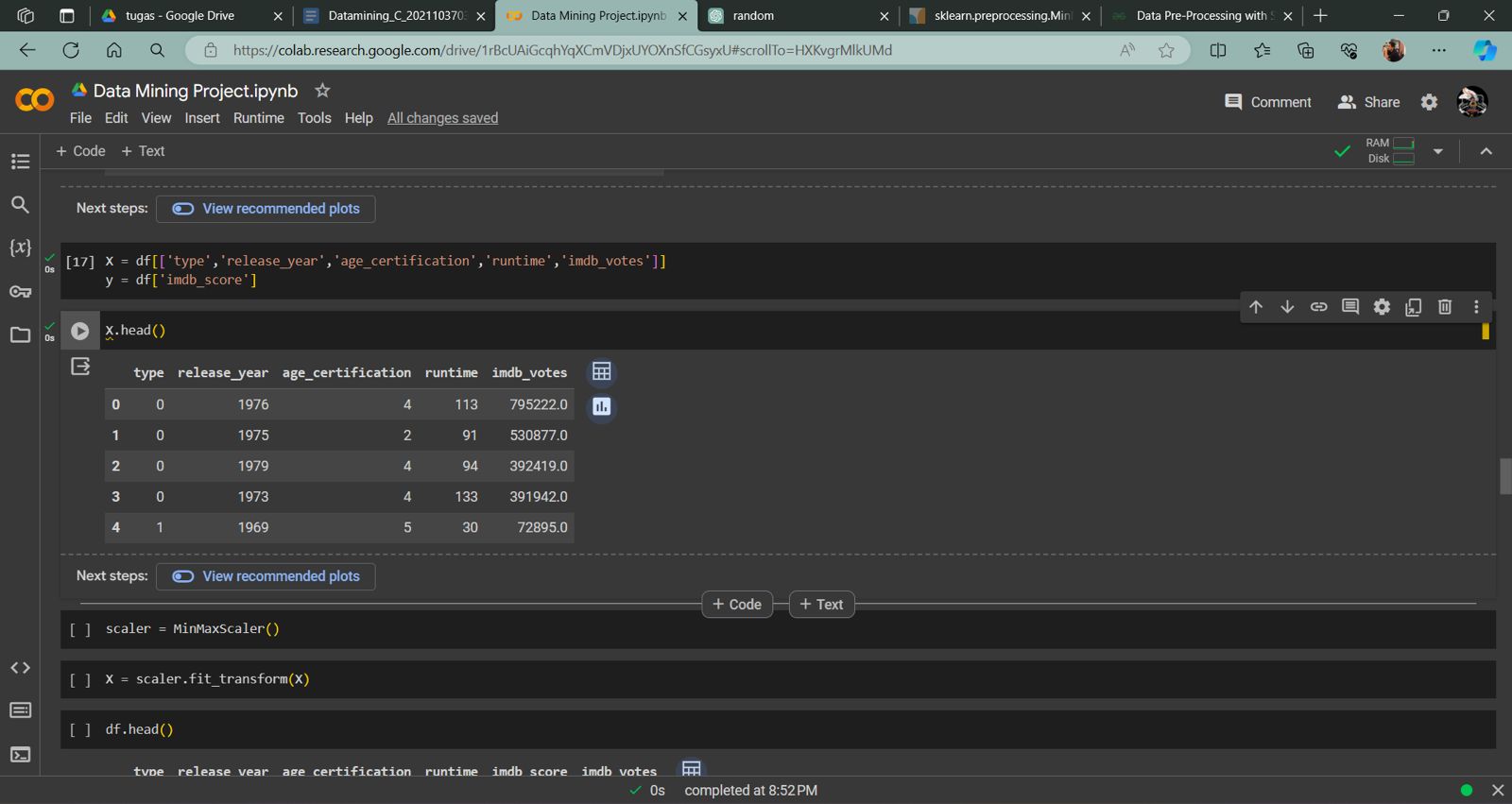


gambar 10 : data setelah dilakukan label encoder

* Data Normalization

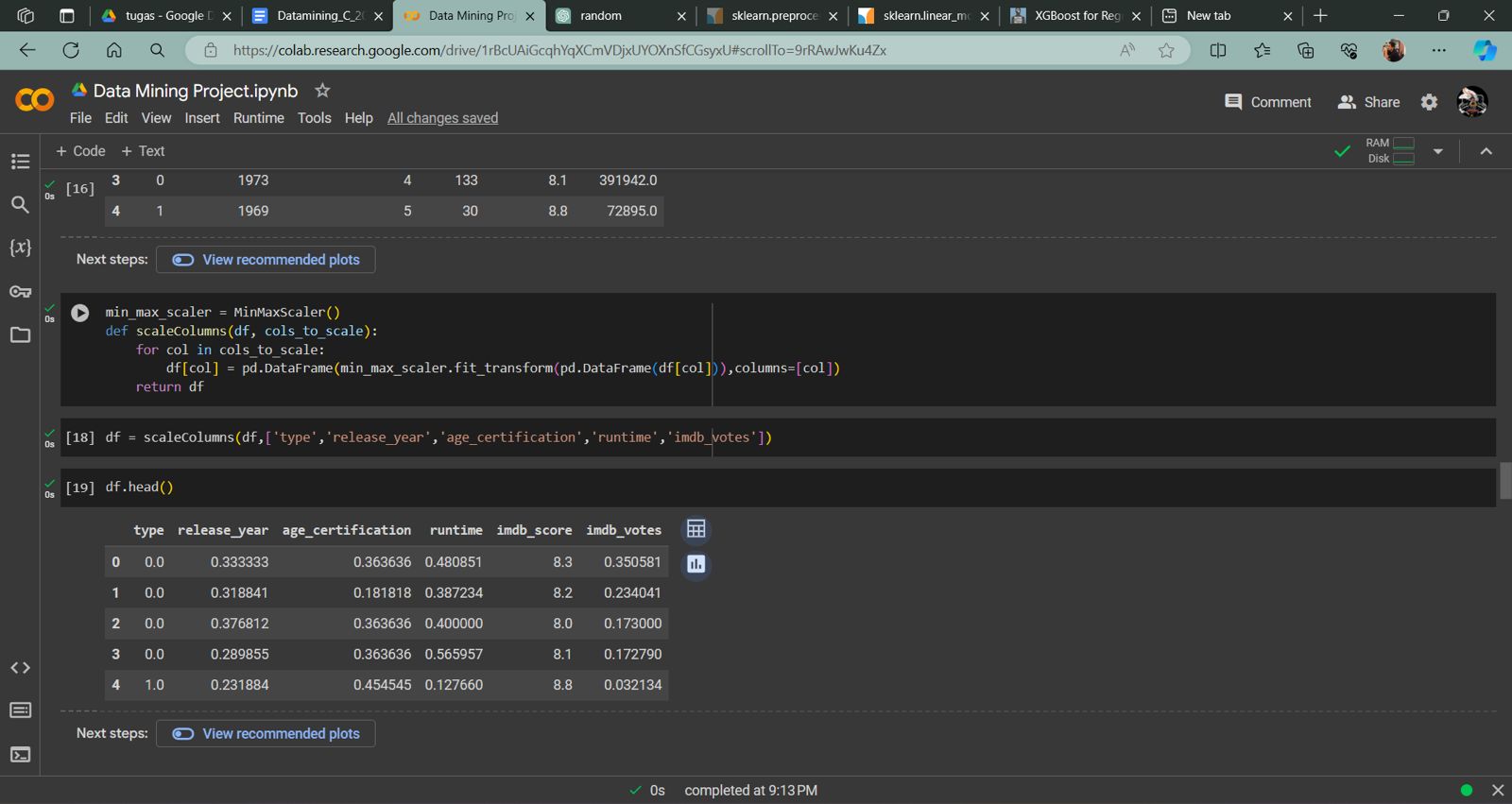
Menggunakan MinMaxScaler, kami standarisasi dependen variabel agar memiliki rentang nilai yang sama

Before :



gambar 11 : data sebelum dinormalisasi

After



gambar 12 : data setelah dinormalisasi

**Tahap 5 (poin: 25)**: Data Mining

Kami menggunakan algoritma Linear Regression untuk memodelkan hubungan antara satu variabel dependent dan independent. Alasan kami mengambil data mining task ini adalah kami ingin melihat variabel mana yang memiliki hubungan erat dengan variabel target. Pertama, kami menentukan target dan fitur. Setelah fitur dan target data telah ditetapkan kami melakukan data splitting menjadi X\_train, X\_test, y\_train, dan y\_test dengan perbandingan 80:20.

Selain menerapkan algoritma regresi pada dataset kami juga membuat tabel korelasi untuk menampilkan hubungan korelasi antara masing-masing atribut pada dataset. Setelah selesai melakukan pemodelan kami melakukan uji coba dengan menggunakan data test yang telah ada hasil dari data splitting sebelumnya. setelah itu kami mencoba menghitung matriks evaluasi kinerja model regresi yang telah dilatih dengan beberapa jenis pengujian. ami menggunakan beberapa algoritma regresi yang tersedia seperti Linear Regression, Random Forest Regressor, Extreme Gradient Boosting Regressor, dan Support Vector Regression. Berikut kami sertakan rumus matematika untuk Algoritma yang gunakan

1. Linear Regression

rumus

Penjelasan :

* y adalah vektor prediksi nilai target.
* adalah matriks fitur input.
* adalah vektor bobot atau koefisien.
* adalah bias atau intercept.

1. Random Forest Regressor

Rumus

penjelasan :

* : Hasil regresi akhir dari Random Forest Regressor untuk input tertentu.
* B : Jumlah total pohon keputusan yang digunakan dalam Random Forest.
* b : Indeks yang menunjukkan urutan pohon keputusan dalam Random Forest.
* : Pohon keputusan ke-bbb dalam Random Forest.
* x : Vektor fitur input yang digunakan untuk membuat prediksi.

1. Extreme Gradient Boosting Regressor

rumus

Di sini:

* adalah prediksi akhir untuk sampel iii pada iterasi t.
* adalah prediksi pada iterasi sebelumnya.
* η adalah learning rate, suatu nilai skalar yang mengontrol kontribusi setiap pohon.
* adalah prediksi dari pohon keputusan ke-t untuk sampel i.

1. Support Vector Regression

rumus

penjelasan :

* adalah prediksi akhir dari model SVR untuk input tertentu x\mathbf{x}x.
* adalah vektor bobot yang menentukan orientasi hyperplane.
* adalah vektor fitur input.
* adalah bias atau offset dari hyperplane.

Setelah menggunakan algoritma regresi yang telah disebutkan diatas saya juga menggunakan metrik yaitu Root Mean Squared Error(RMSE) dan Mean Absolute Error(MAE) untuk menjadi bahan evaluasi model yang saya buat. Metrik ini juga memiliki rumus matematika nya sendiri yaitu :

1. Root Mean Squared Error(RMSE):

penjelasan :

* n adalah jumlah total sampel.
* adalah nilai target sebenarnya untuk sampel ke i.
* ​ adalah nilai prediksi untuk sampel ke ii.

1. Mean Absolute Error(MAE):

MAE =

Penjelasan :

* n adalah jumlah total sampel.
* adalah nilai target sebenarnya untuk sampel ke i.
* ​ adalah nilai prediksi untuk sampel ke ii.

Tabel perbandingan Menggunakan data original

| Data | Algoritma/ Seleksi Fitur | Metriks | Linear Regression | Random Forest | XGBoost | Support Vector Regression |
| --- | --- | --- | --- | --- | --- | --- |
| original | original | RMSE | 1,04 | 1,13 | 1,02 | 1,05 |
| MAE | 7,97 | 8,03 | 7,85 | 7,94 |
| chi square | RMSE | 1,42 | 1,67 | 1,38 | 1,42 |
|  | MAE | 1,10 | 1,27 | 1,06 | 1,10 |
| SFS forward | RMSE | 1,04 | 1,06 | 1,02 | 1,05 |
|  | MAE | 0,80 | 0,82 | 0,79 | 0,79 |
| SFS backward | RMSE | 1,04 | 1,09 | 1,02 | 1,05 |
|  | MAE | 0,80 | 0,83 | 0,79 | 0,79 |
| RFE | RMSE | 1,05 | 1,09 | 1,03 | 1,05 |
|  | MAE | 0,81 | 0,83 | 0,80 | 0,80 |
| Resampled | original | RMSE | 1,04 | 8,36 | 1,04 | 1,37 |
|  | MAE | 1,12 | 5,38 | 1,04 | 1,10 |
| chi square | RMSE | 1,74 | 1,42 | 1,60 | 1,78 |
|  | MAE | 1,44 | 0,99 | 1,32 | 1,42 |
| SFS forward | RMSE | 1,38 | 0,85 | 1,24 | 1,38 |
|  | MAE | 1,14 | 0,55 | 1,04 | 1,12 |
| SFS backward | RMSE | 1,38 | 0,85 | 1,24 | 1,38 |
|  | MAE | 1,14 | 0,55 | 1,04 | 1,12 |
| RFE | RMSE | 1,37 | 0,86 | 1,26 | 1,37 |
|  | MAE | 1,13 | 0,56 | 1,05 | 1,11 |
| Hasil Paling kecil | | RMSE | 1,04 | 0,85 | 1,02 | 1,05 |
| MAE | 0,8 | 0,55 | 0,79 | 0,79 |

tabel 1 : tabel perbandingan hasil algoritma dengan data, dan penerapan seleksi fitur

Pada tabel perbandingan di atas dapat kita lihat perbandingan dari RMSE dan MAE menggunakan data, algoritma dan seleksi fitur yang berbeda-beda. RMSE dari XGboost dengan data original tanpa seleksi fitur mendapatkan point yang paling rendah. Sedangkan untuk MAE, point paling rendah ada diperoleh oleh Random forest dengan data resampled dan seleksi fitur SFS backward. Mehilahat dari rendah nya RMSE dari Xgboost kami menampilkan hasil feature importance atau tingkat kepentingan suatu variabel independen terhadap variabel independen

Extreme Gradient Boosting Regressor dengan original data tanpa seleksi fitur :

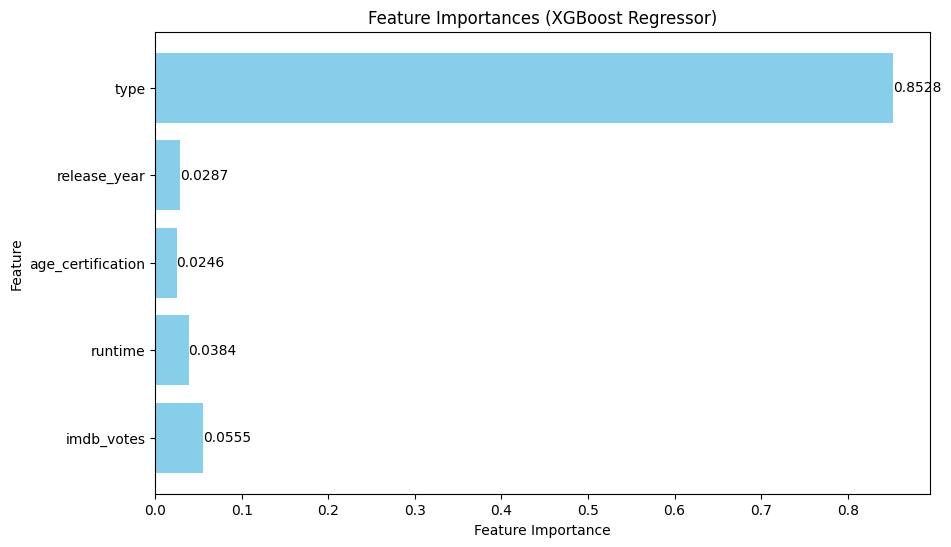
feature\_importances

release\_year: 0.0287

age\_certification: 0.0246

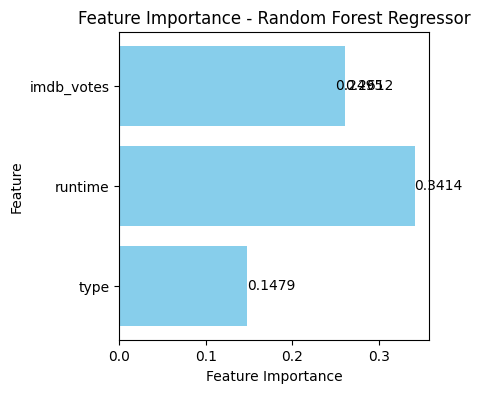
runtime: 0.0384

imdb\_votes: 0.0555



gambar 13 : diagram nilai kepentingan fitur XGBoost

Random Forest Regressor dengan resampled data dan SFS backward



gambar 14 :diagram nilai kepentingan fitur Random forest

Pada gambar 13 dapat kita lihat bahwa algoritma XGBoost menilai tipe dari suatu tontonan memiliki nilai pengaruh atau nilai kepentingan yang paling tinggi. sementara pada gambar 14, yang memperoleh nilai paling tinggi adalah *runtime* atau durasi dari tontonan.

**Tahap 6 (poin: 20)**: Knowledge Interpretation

Berdasarkan hasil dari implementasi algoritma regresi serta memperhatikan tabel korelasi, dapat diperhatikan pada bagian koefisiensi dan feature importances bahwa atribut runtime dan imdb\_votes memiliki nilai yang lebih tinggi daripada atribut lain. Kami menyimpulkan bahwa dua atribut ini lah yang memiliki pengaruh yang lebih besar daripada atribut lain. Akan tetapi perlu di ingat masih ada beberapa atribut lain yang harus nya memiliki pengaruh pada popularitas suatu film atau acara TV seperti trailer, judul dan deskripsi nya

1. Faktor Lain yang Mempengaruhi Popularitas Film dan Acara TV

Walaupun hasil analisis regresi menunjukkan bahwa runtime dan imdb\_votes memiliki pengaruh besar terhadap popularitas film atau acara TV, penting untuk diingat bahwa masih terdapat faktor lain yang tidak dipertimbangkan dalam model. Faktor-faktor ini dapat dikategorikan menjadi dua:

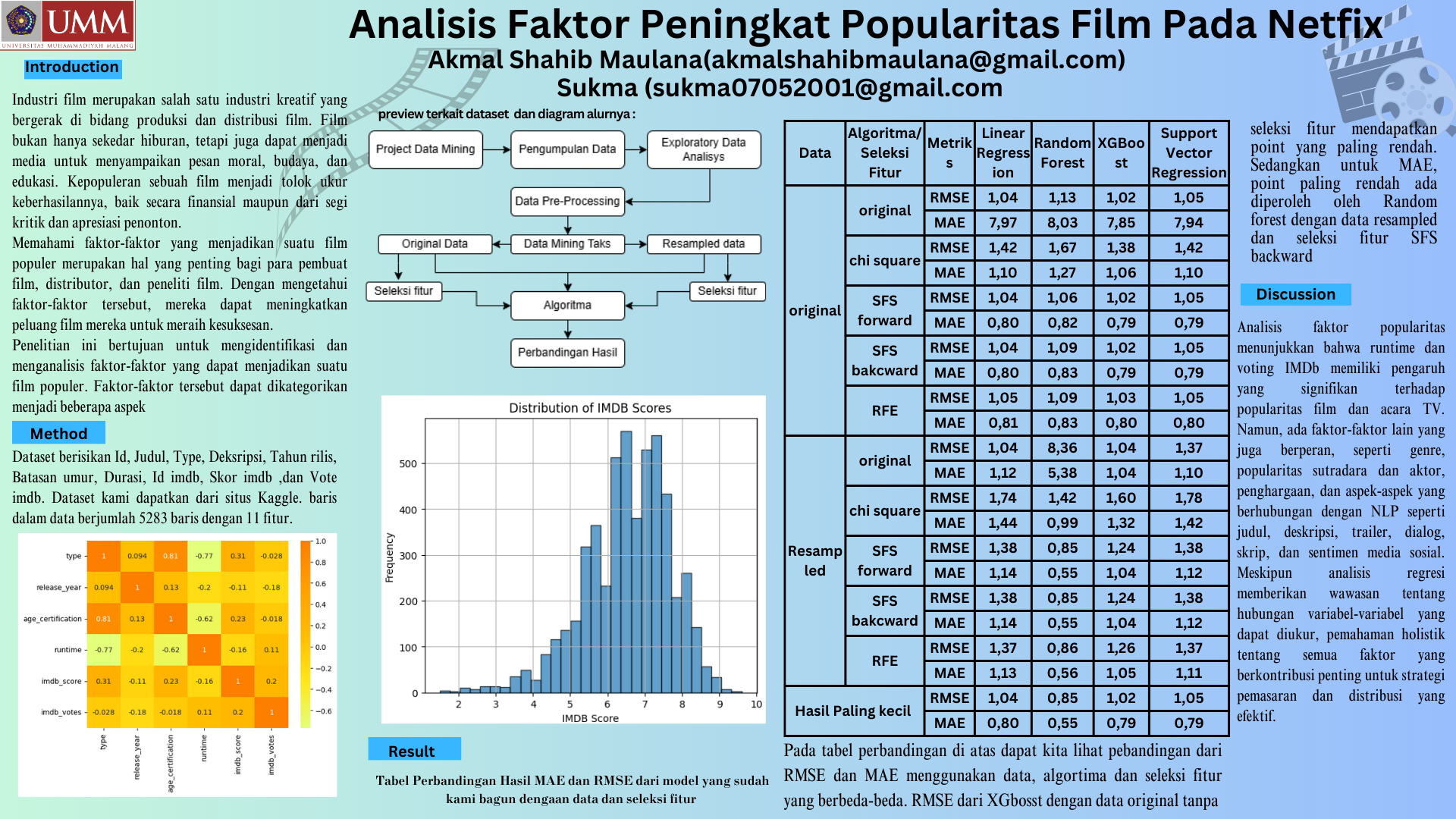
1. Faktor yang Dapat Dikategorikan:
   1. Genre: Genre film atau acara TV dapat significantly mempengaruhi popularitasnya. Contohnya, film action dan komedi umumnya lebih populer daripada film drama atau dokumenter.
   2. Sutradara dan Aktor: Popularitas sutradara dan aktor terkenal dapat menarik penonton dan meningkatkan popularitas film atau acara TV.
   3. Penghargaan: Kemenangan di festival film bergengsi atau penghargaan lain dapat meningkatkan popularitas film atau acara TV.
2. Faktor yang Berkaitan dengan NLP (Natural Language Processing):
   1. Judul dan Deskripsi: Judul dan deskripsi yang menarik dan informatif dapat menarik perhatian penonton dan mendorong mereka untuk menonton film atau acara TV.
   2. Trailer: Trailer yang menarik dan informatif dapat memberikan gambaran tentang film atau acara TV dan menarik minat penonton.
   3. Dialog dan Naskah: Kualitas dialog dan naskah dapat mempengaruhi pengalaman menonton dan kepuasan penonton.
   4. Sentimen Media Sosial: Sentimen positif di media sosial tentang film atau acara TV dapat meningkatkan popularitasnya.

Kesimpulan:

Meskipun analisis regresi memberikan wawasan tentang hubungan antara variabel yang dapat diukur secara kuantitatif, penting untuk mempertimbangkan faktor-faktor lain yang mungkin memiliki pengaruh signifikan terhadap popularitas film atau acara TV. Memahami faktor-faktor ini secara holistik dapat membantu dalam mengembangkan strategi pemasaran dan distribusi yang lebih efektif untuk meningkatkan popularitas film atau acara TV.

**Tahap 7 (poin: 15)**: Reporting

* Simple academic Poster.



[Link canva](https://www.canva.com/design/DAGHKLWy3q8/5XbiqeEZG1TEGqMhBek2wQ/edit?utm_content=DAGHKLWy3q8&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton)

Jupyter Notebook (Python)

[Data Mining Project.ipynb - Colab (google.com)](https://colab.research.google.com/drive/1rBcUAiGcqhYqXCmVDjxUYOXnSfCGsyxU)