FINAL Project

Prediksi Penyakit Asma

- By Eddy Refianto
- 9 october, 2023

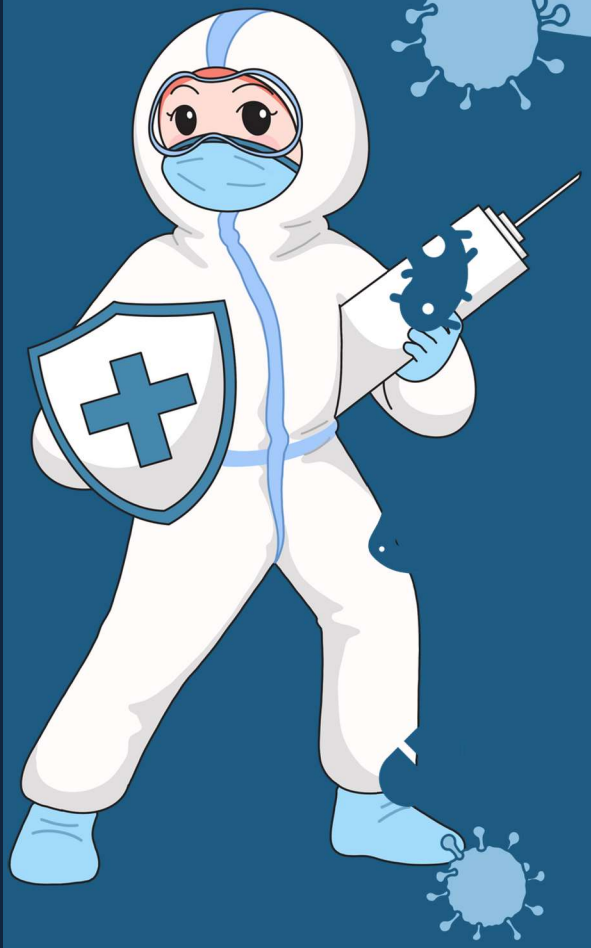
Contents

Introduction

- Kami akan melakukan analisis data eksplorasi pada kumpulan data Kaggle [Prediksi Penyakit Asma] (https://www.kaggle.com/datasets/deepayanthakur/asthma-disease-prediction) by Deepayan Thakur.
- Dataset ini merupakan kumpulan komprehensif berbagai gejala dan faktor dari pasien dengan atau tanpa asma.
- Tujuan dari proyek ini adalah untuk membangun model pembelajaran mesin yang dapat memprediksi apakah seorang pasien menderita asma atau tidak berdasarkan gejala dan faktornya, dan jika ya, seberapa parah asmanya.

Dataset

```
RangeIndex: 316800 entries, 0 to 316799
Data columns (total 19 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   Tiredness            316800 non-null  int64
 1   Dry-Cough            316800 non-null  int64
 2   Difficulty-in-Breathing  316800 non-null  int64
 3   Sore-Throat          316800 non-null  int64
 4   None_Sympton         316800 non-null  int64
 5   Pains                316800 non-null  int64
 6   Nasal-Congestion     316800 non-null  int64
 7   Runny-Nose           316800 non-null  int64
 8   None_Experiencing    316800 non-null  int64
 9   Age_0-9              316800 non-null  int64
 10  Age_10-19            316800 non-null  int64
 11  Age_20-24            316800 non-null  int64
 12  Age_25-59            316800 non-null  int64
 13  Age_60+              316800 non-null  int64
 14  Gender_Female        316800 non-null  int64
 15  Gender_Male          316800 non-null  int64
 16  Severity_Mild        316800 non-null  int64
 17  Severity_Moderate    316800 non-null  int64
 18  Severity_None        316800 non-null  int64
dtypes: int64(19)
```



Exploratory Data Analysis

Hapus data duplikat

- Menghapus data duplikat agar lebih memudahkan untuk Machine Learningnya tidak terjadi perulangan pembelajaran dengan data yang sama.
- Dari baris data 316 ribu setelah dilalukan Data Cleaning dengan menghapus duplikasi data maka diperoleh 5760 baris data.

Kategori

Membalikkan pengkodean *one-hot* dan mengubah fitur-fitur tersebut menjadi fitur kategorikal untuk Column Severity untuk membantu visualisasi data.
'Severity_Mild',
'Severity_Moderate',
'Severity_None'

Imbalanced

baris data dengan tingkat Severity antara 0,1, dan 2 berbeda maka perlu diseimbangkan memakai undersample.

2880

'Severity_Mild',

1440

'Severity_Moderate'

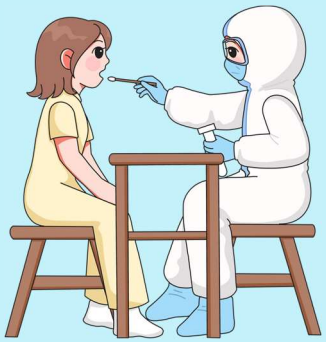
1440

'Severity_None’

Pre-processing

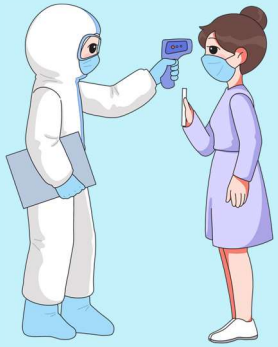
Outlier

Memakai Z-Score



one-hot encoding

Kategori age_mapping = {
'0-9': 0,
'10-19': 1,
'20-24': 2,
'25-59': 3,
'60+': 4
} dan gender_mapping = {
'Male': 0,
'Female': 1
}



Heatmap

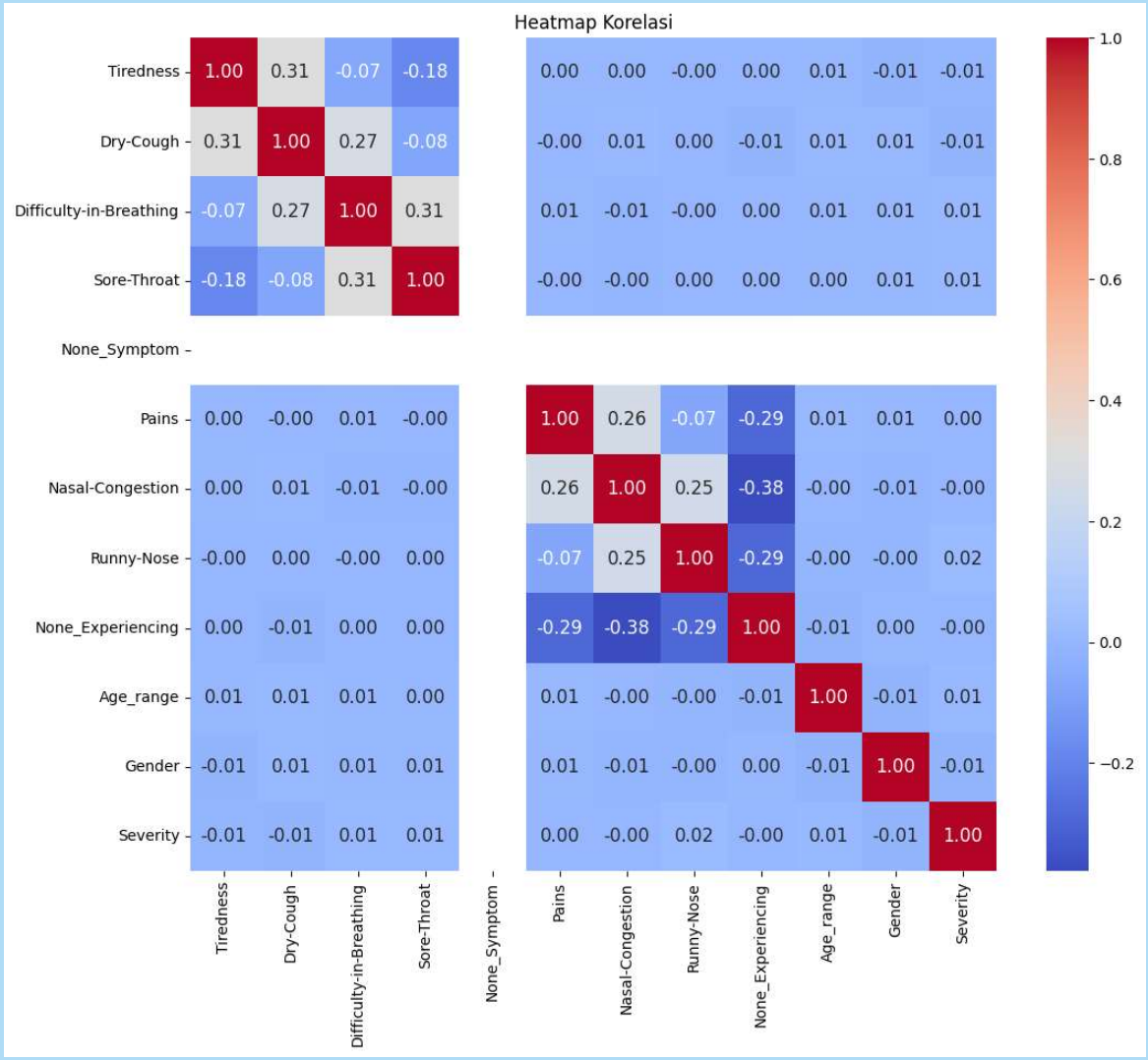
Keterkaitan antar column

Scalar

Menggunakan scaler untuk mengubah data ke skala, tetapi tidak banyak pengaruh maa tidak dipakai.

Heat Map

- Karena Column None_Sympton berisi Nan dan Column None Experincing tidak punya pengaruh maka 2 variable ini tidak dimasukkan dalam permodelan.



Matrik Kebingungan

- Baris pertama menggambarkan label "Mild" pada data pengujian:

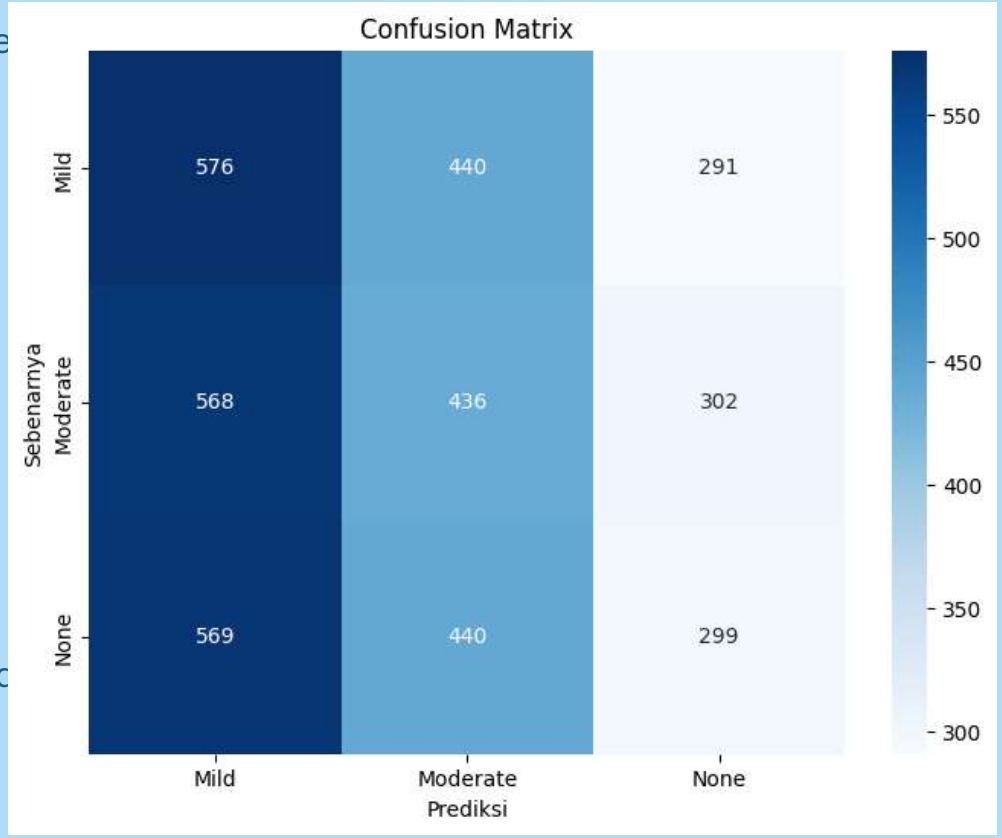
- 576 data dengan label "Mild" diprediksi dengan benar sebagai "Mild" (True Positive untuk "Mild").
- 440 data dengan label "Mild" salah diprediksi sebagai "Moderate" (False Negative untuk "Mild").
- 291 data dengan label "Mild" salah diprediksi sebagai "None" (False Negative untuk "Mild").

- Baris kedua menggambarkan label "Moderate" pada data pengujian:

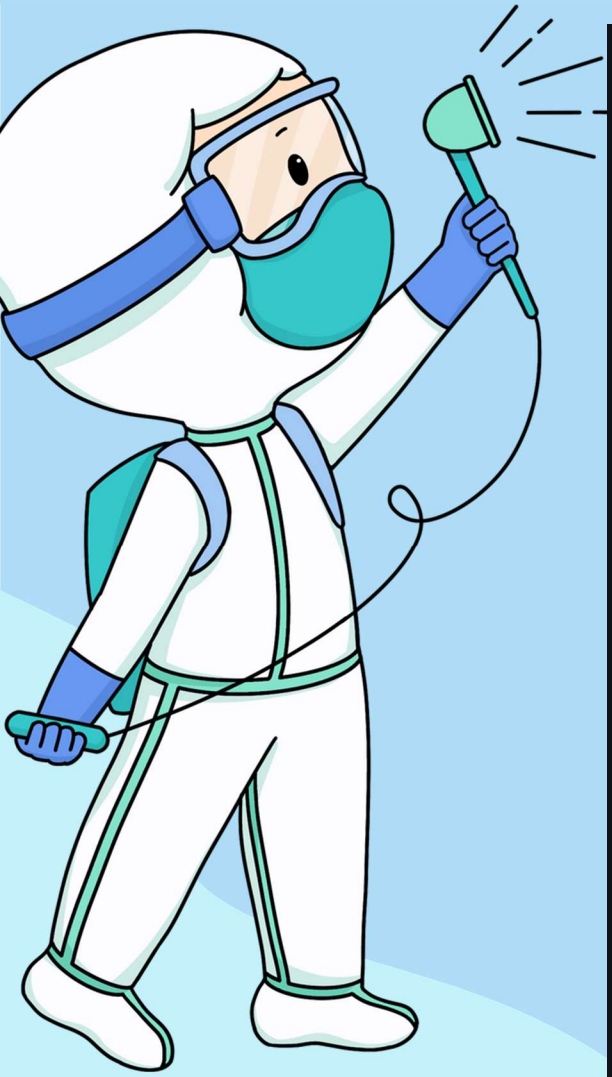
- 568 data dengan label "Moderate" salah diprediksi sebagai "Mild" (False Negative untuk "Moderate").
- 436 data dengan label "Moderate" diprediksi dengan benar sebagai "Moderate" (True Positive untuk "Moderate").
- 302 data dengan label "Moderate" salah diprediksi sebagai "None" (False Negative untuk "Moderate").

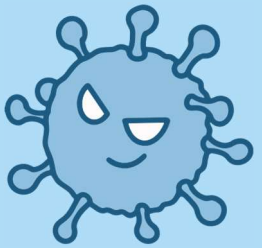
- Baris ketiga menggambarkan label "None" pada data pengujian:

- 569 data dengan label "None" salah diprediksi sebagai "Mild" (False Negative untuk "None").
- 440 data dengan label "None" salah diprediksi sebagai "Moderate" (False Negative untuk "None").
- 299 data dengan label "None" diprediksi dengan benar sebagai "None" (True Positive untuk "None").



Deployment





Please Insert The Informations

Tiredness

Yes

Dry Cough

No

Difficulty in Breathing

Yes

Sore Throat

Yes

Pains

No

Nasal Congestion

No

Runny Nose

Yes

Age Range

60+

Gender

Male

Predict

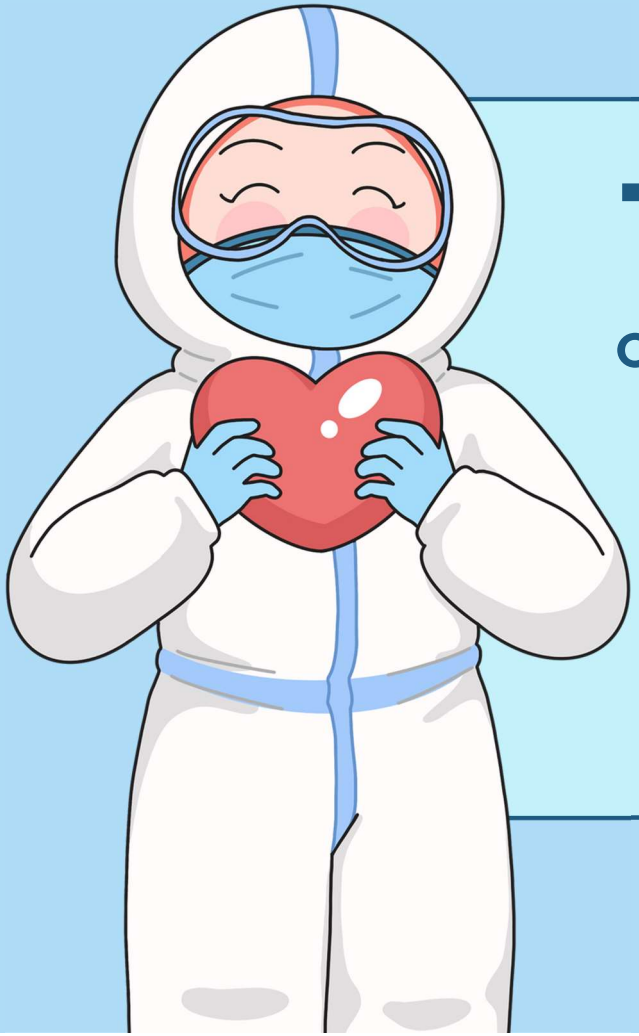
Severity: Moderate

Project Conclusions

Dari Confusion matrix tersebut maka diambil data yang sesuai saja yaitu data Actual yang sama dengan Predicted menjadi dataset baru.

Maka dilakukan manipulasi data dengan menggunakan 99% data test yang sudah diuji akurasinya mencapai 0.98.

Kemudian dilakukan modeling dan deploy menggunakan streamlit.



THANK YOU!

Contact us if you have questions