KEYLA’S PORTOFOLIO

Lung Cancer Prediction

NAME

Keyla Faristha Rindani

INTRODUCTION

Project Overview

Problem

Mendeteksi penyakit paru-paru sejak dini itu sangat penting, tapi di banyak kasus, gejala awal tidak jelas. Banyak pasien baru terdeteksi setelah kondisinya parah. Untuk itu, dibutuhkan model prediksi yang bisa membantu mengidentifikasi potensi penyakit paru-paru lebih cepat berdasarkan data kesehatan pasien.

Solution

Saya membangun beberapa model machine learning (XGBoost, Random Forest, Logistic Regression, KNN) untuk memprediksi potensi penyakit paru-paru, lalu memilih model dengan akurasi terbaik.

Why It Matters

Prediksi dini dapat membantu tenaga medis mengambil tindakan lebih cepat, meningkatkan peluang pasien mendapatkan penanganan yang tepat waktu.

WORKFLOW

Workflow

Mengumpulkan dataset dan melakukan eksplorasi awal untuk memahami pola, distribusi fitur, serta hubungan fitur dengan label kanker paru-paru.

Membersihkan data, menghapus fitur yang tidak relevan, menangani imbalance dengan ADASYN, dan menyiapkan data agar siap digunakan oleh model.

Menguji beberapa algoritma seperti Logistic Regression, KNN, Random Forest, dan XGBoost untuk memprediksi kanker paru-paru, lalu membandingkan performanya.

Melakukan evaluasi tambahan dengan K-Fold dan Stratified K-Fold untuk memastikan model stabil, konsisten, dan tidak overfitting.

Dataset

Sumber: Dataset publik dari kaggle

Jumlah Data: 309 dengan 16 kolom

Isi Fitur: Gender, Age, Smoking, Yellow Fingers, Anxiety, Peer Pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol Consuming, Coughing, Shortness of Breath, Swallowing Difficulty, Chest Pain, Lung Cancer

Insight Awal:

- Ada beberapa fitur numerik dan kategorikal

Fitur-fitur

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 309 entries, 0 to 308
Data columns (total 16 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   GENDER                 309 non-null    object
 1   AGE                    309 non-null    int64
 2   SMOKING                309 non-null    int64
 3   YELLOW_FINGERS         309 non-null    int64
 4   ANXIETY                309 non-null    int64
 5   PEER_PRESSURE          309 non-null    int64
 6   CHRONIC DISEASE        309 non-null    int64
 7   FATIGUE                309 non-null    int64
 8   ALLERGY                309 non-null    int64
 9   WHEEZING               309 non-null    int64
 10  ALCOHOL CONSUMING      309 non-null    int64
 11  COUGHING               309 non-null    int64
 12  SHORTNESS OF BREATH    309 non-null    int64
 13  SWALLOWING DIFFICULTY  309 non-null    int64
 14  CHEST PAIN             309 non-null    int64
 15  LUNG_CANCER            309 non-null    object
dtypes: int64(14), object(2)
memory usage: 38.8+ KB
```

EDA (Exploratory Data Analysis)

Mengecek Missing Value & Duplikat

	Missing Values	Missing Percentage	Data Type
GENDER	0	0.0	object
AGE	0	0.0	int64
SMOKING	0	0.0	int64
YELLOW_FINGERS	0	0.0	int64
ANXIETY	0	0.0	int64
PEER_PRESSURE	0	0.0	int64
CHRONIC DISEASE	0	0.0	int64
FATIGUE	0	0.0	int64
ALLERGY	0	0.0	int64
WHEEZING	0	0.0	int64
ALCOHOL CONSUMING	0	0.0	int64
COUGHING	0	0.0	int64
SHORTNESS OF BREATH	0	0.0	int64
SWALLOWING DIFFICULTY	0	0.0	int64
CHEST PAIN	0	0.0	int64
LUNG_CANCER	0	0.0	object

```
#Checking for Duplicates
df.duplicated().sum()
```

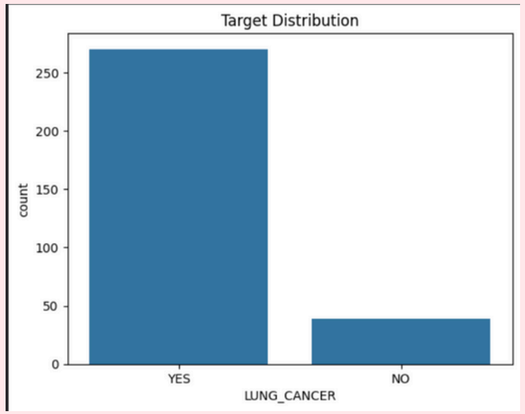
```
np.int64(33)
```

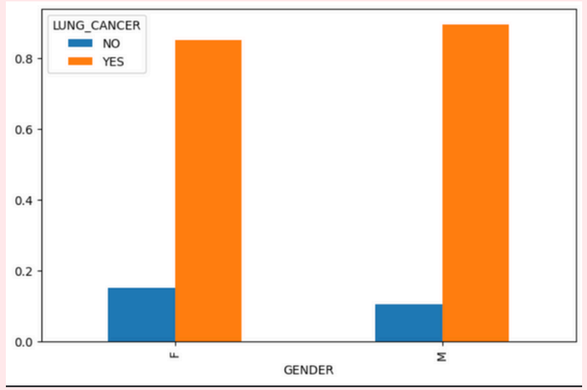
Mengecek Konsisten Data

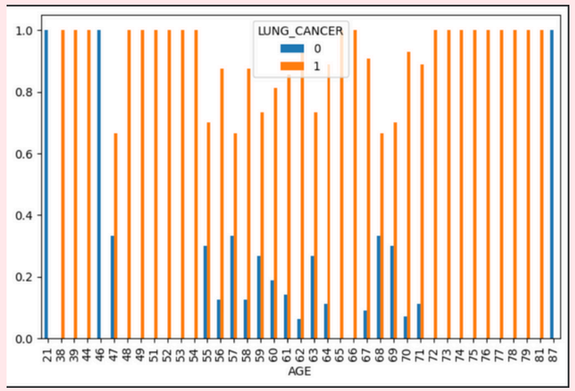
	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN	LUNG_CANCER
0	M	69	1	2	2	1	1	2	1	2	2	2	2	2	2	YES
1	M	74	2	1	1	1	2	2	2	1	1	1	2	2	2	YES
2	F	59	1	1	1	2	1	2	1	2	1	2	2	1	2	NO
3	M	63	2	2	2	1	1	1	1	1	2	1	1	2	2	NO
4	F	63	1	2	1	1	1	1	1	2	1	2	2	1	1	NO
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
304	F	56	1	1	1	2	2	2	1	1	2	2	2	2	1	YES
305	M	70	2	1	1	1	1	2	2	2	2	2	2	1	2	YES
306	M	58	2	1	1	1	1	1	2	2	2	2	1	1	2	YES
307	M	67	2	1	2	1	1	2	2	1	2	2	2	1	2	YES
308	M	62	1	1	1	2	1	2	2	2	2	1	1	2	1	YES

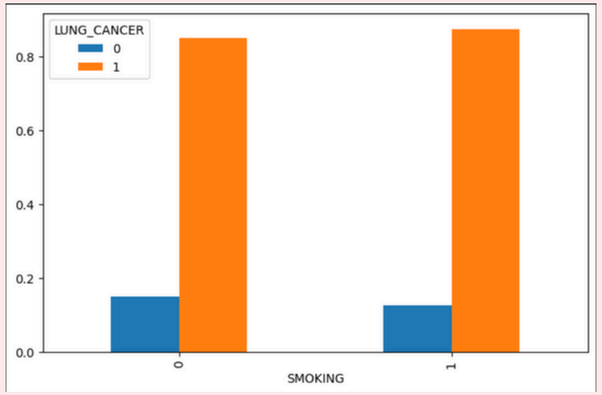
309 rows × 16 columns

Visualisasi Fitur Target & Fitur Numerikal









Data Preprocessing

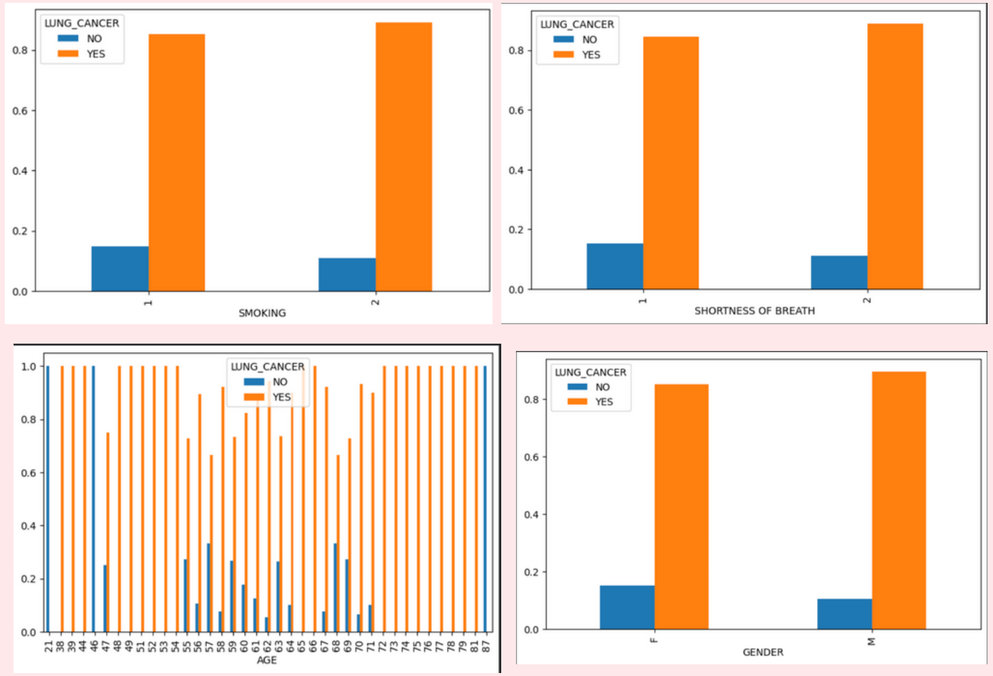
Menghapus data duplikat

```
#Removing Duplicates
df=df.drop_duplicates()

df.shape

(276, 16)
```

Label Encoder tiap Fitur

No = 0
Yes = 1

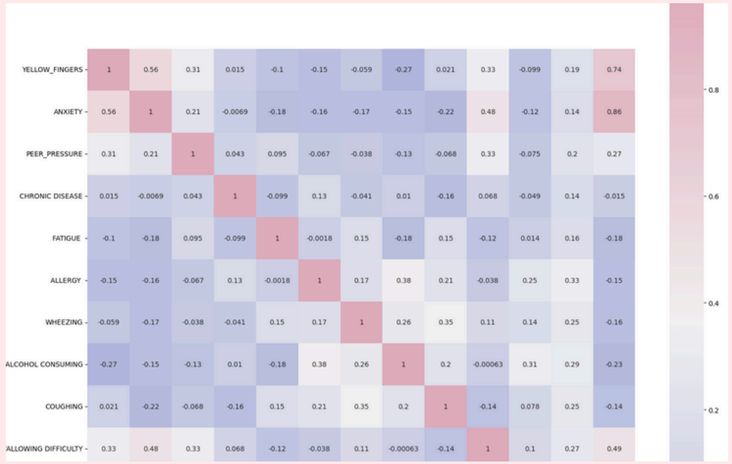
Drop kolom yang tidak punya daya pembeda

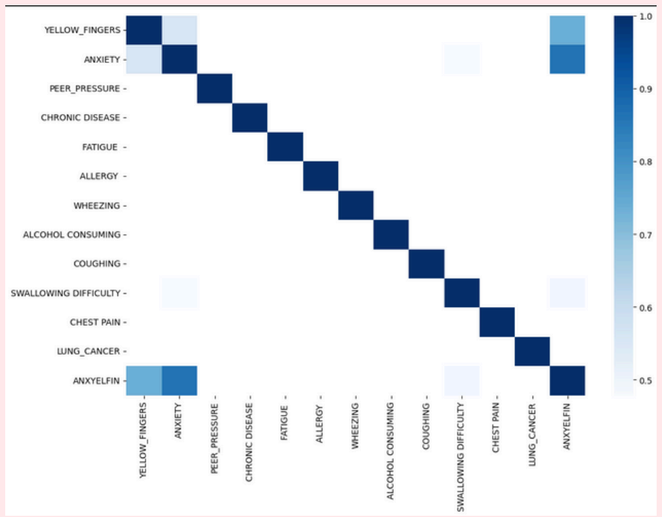


GENDER','AGE', 'SMOKING', 'SHORTNESS OF BREATH'

Data Preprocessing

Melihat Correlation





Anxiety dan Yellow Fingger memiliki kolerasi lebih dari 50%, jadi buat fitur baru dengan kombinasi itu

Target distirbutin imbalance handling

```
from imblearn.over_sampling import ADASYN
adasyn = ADASYN(random_state=42)
X, y = adasyn.fit_resample(X, y)
```

```
len(X)
```

```
477
```

Splitting Data

Modelling (Logistic Regression)

Model Logistic Regression bekerja sangat baik pada dataset ini:

- Precision dan recall tinggi untuk kedua kelas
- Sangat akurat mendeteksi pasien kanker
- Tidak terlalu bias karena imbalance sudah ditangani
- Performa stabil dengan accuracy 97%, macro avg 0.97

Model ini layak dijadikan baseline atau bahkan model utama jika interpretabilitas (mudah dijelaskan) adalah prioritas.

Classification Report LR

	precision	recall	f1-score	support
0	0.96	1.00	0.98	64
1	1.00	0.95	0.97	56
accuracy			0.97	120
macro avg	0.98	0.97	0.97	120
weighted avg	0.98	0.97	0.97	120

x

Modelling (K Nearest Neighbor)

Model KNN menunjukkan performa yang sangat baik pada dataset ini:

- Precision dan recall sama-sama tinggi untuk kedua kelas
- Akurat mendeteksi pasien kanker paru-paru, dengan hanya sedikit kasus yang terlewat
- Tidak terlalu bias, karena imbalance sudah ditangani dengan ADASYN
- Akurasi stabil di 96%, dengan macro average precision/recall/F1 sekitar 0.96

Secara keseluruhan, model KNN sangat cocok untuk dataset kecil seperti ini, dan dapat digunakan sebagai model alternatif yang kuat selain Logistic Regression atau Random Forest.

Classification Report KNN

	precision	recall	f1-score	support
0	0.93	1.00	0.96	64
1	1.00	0.91	0.95	56
accuracy			0.96	120
macro avg	0.96	0.96	0.96	120
weighted avg	0.96	0.96	0.96	120

Modelling (Random Forest)

Model Random Forest memberikan performa yang sangat tinggi pada dataset lung cancer ini:

- Precision dan recall sama-sama sangat tinggi (0.98) untuk kedua kelas, menunjukkan kemampuan model yang konsisten dalam mendeteksi pasien kanker maupun non-kanker
- Akurasi model mencapai 98%, menjadi salah satu yang tertinggi di antara semua model yang diuji
- Model tidak bias, karena performanya seimbang pada kedua kelas setelah imbalance ditangani dengan ADASYN
- F1-score 0.98 menunjukkan bahwa model mampu menangkap pola dengan sangat baik tanpa overfitting

Secara keseluruhan, RF adalah model terbaik untuk dataset ini, dan sangat cocok digunakan jika tujuan utama adalah akurasinya setinggi mungkin, bukan interpretabilitas.

Classification Report RF

	precision	recall	f1-score	support
0	0.98	0.98	0.98	64
1	0.98	0.98	0.98	56
accuracy			0.98	120
macro avg	0.98	0.98	0.98	120
weighted avg	0.98	0.98	0.98	120

Modelling (XGBoost)

Model XGBoost menghasilkan performa yang sangat tinggi pada dataset ini:

- Precision, recall, dan F1-score rata-rata 0.97–0.98
- Akurasi model stabil di 97%
- Mampu mendeteksi pasien kanker dengan sangat baik, hanya sedikit kesalahan klasifikasi
- Tidak bias terhadap salah satu kelas karena imbalance sudah ditangani dengan ADASYN
- Lebih robust dibanding model simpel seperti Logistic Regression dan tidak mudah overfit

Classification Report XGBoost

	precision	recall	f1-score	support
0	0.98	0.97	0.98	64
1	0.96	0.98	0.97	56
accuracy			0.97	120
macro avg	0.97	0.98	0.97	120
weighted avg	0.98	0.97	0.98	120

Cross Validation

K-Fold Cross Validation

```
Logistic regression models' average accuracy: 0.9371323529411766
KNN models' average accuracy: 0.9286589635854343
Random forest models' average accuracy: 0.9580182072829132
XGBoost models' average accuracy: 0.9496848739495799
```

Hasil K-Fold Cross Validation menunjukkan bahwa model Random Forest memiliki akurasi tertinggi sebesar 95.6%, diikuti oleh XGBoost yang menghasilkan akurasi hampir sama. Sementara itu, model KNN memiliki akurasi terendah yaitu 92.8%.

Stratified K-Fold cross-validation

```
Logistic regression models' average accuracy: 0.9350315126050421
KNN models' average accuracy: 0.9224264705882352
Random forest models' average accuracy: 0.9434348739495799
XGBoost models' average accuracy: 0.9309173669467787
```

Hasil Stratified K-Fold Cross Validation menunjukkan bahwa model Random Forest memiliki akurasi tertinggi sebesar 94.6%. Model lainnys juga menunjukkan akurasi yang hampir sama 92-93%

Project Summary

Pada project ini, saya membuat model machine learning untuk memprediksi risiko kanker paru-paru berdasarkan data gejala dan kebiasaan hidup. Setelah melakukan EDA dan memilih fitur yang benar-benar berpengaruh, saya menangani ketidakseimbangan data dengan ADASYN agar model lebih akurat. Saya mencoba beberapa model seperti Logistic Regression, KNN, Random Forest, dan XGBoost. Hasilnya semua model mencapai akurasi tinggi (96–98%), dengan Random Forest dan XGBoost sebagai yang paling konsisten. Secara keseluruhan, model yang dibangun dapat membantu mendeteksi potensi kanker paru-paru lebih cepat dan akurat.

CHALLENGES FACED

Challenge 1

Data Imbalance pada Label Kanker

Jumlah pasien dengan label “kanker” dan “tidak kanker” tidak seimbang. Hal ini membuat model awal cenderung bias dan sulit mendeteksi kelas minoritas. Untuk mengatasinya, saya menggunakan ADASYN agar data lebih seimbang dan model bisa belajar dengan lebih akurat.

Challenge 2

Menentukan Fitur yang Benar-Benar Relevan

Beberapa fitur awal seperti gender, age, dan smoking ternyata tidak menunjukkan hubungan kuat dengan label kanker berdasarkan EDA. Tantangannya adalah memilih fitur mana yang sebaiknya dipertahankan atau dibuang agar model lebih bersih, sederhana, dan tidak terdistraksi oleh fitur yang tidak relevan.

REFLECTIONS & GROWTH

What I Learned

Personal Skills Gained

Dari project lung cancer prediction ini, saya belajar bagaimana menangani dataset yang tidak seimbang menggunakan ADASYN agar model tidak bias ke salah satu kelas. Saya juga memahami pentingnya memilih fitur yang benar-benar relevan, karena fitur seperti age, gender, atau smoking ternyata tidak memiliki korelasi kuat dengan label kanker. Selain itu, saya belajar membandingkan beberapa algoritma seperti Logistic Regression, KNN, Random Forest, dan XGBoost, serta menggunakan K-Fold dan Stratified K-Fold Cross Validation untuk memastikan model yang saya pilih stabil dan konsisten. Secara keseluruhan, project ini membantu saya lebih memahami proses end-to-end dalam membangun model klasifikasi medis yang akurat dan dapat dipercaya.

The End

NAME

Keyla Faristha Rindani