Kampus Merdeka INDONESIA JAYA | orbit FUTURE ACADEMY | Skills For Future Jobs | Institut Teknologi Telkom Purwokerto Bridging Technology for Humanity

Fayza Apriliza - Broyden

PREDIKSI INDEKS PEMBANGUNAN MANUSIA

menggunakan Decision Tree, Random Forest, Adaboost, XGBoost, dan Teknik Stacking.

PROJECT INI BERTUJUAN UNTUK MEMPREDIKSI STATUS IPM SESEORANG BERDASARKAN HARAPAN LAMA SEKOLAH, PENGELUARAN PERKAPITA, RERATA LAMA SEKOLAH, DAN USIA HARAPAN HIDUP.

1. Bagaimana persebaran dari setiap variabel?
2. Apakah terdapat korelasi atau hubungan antara variabel dependent dan independen?
3. Mana model yang paling baik dan berapa skor akurasinya?

DATA

PROJECT INI MENGGUNAKAN DATA INDEKS PEMBANGUNAN MANUSIA SEBANYAK 2196 BARIS

Harapan lama sekolah

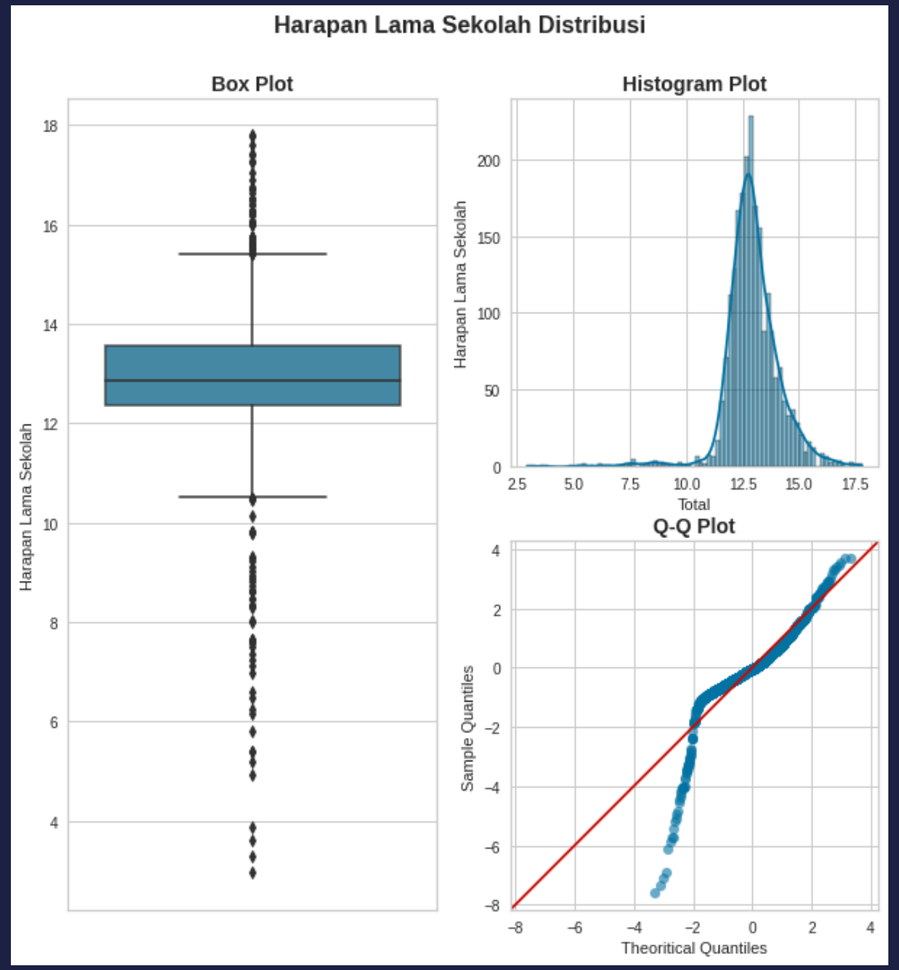
Pengeluaran perkapita

Rerata lama sekolah

Usia harapan hidup

IPM

DISTRIBUSI DATA



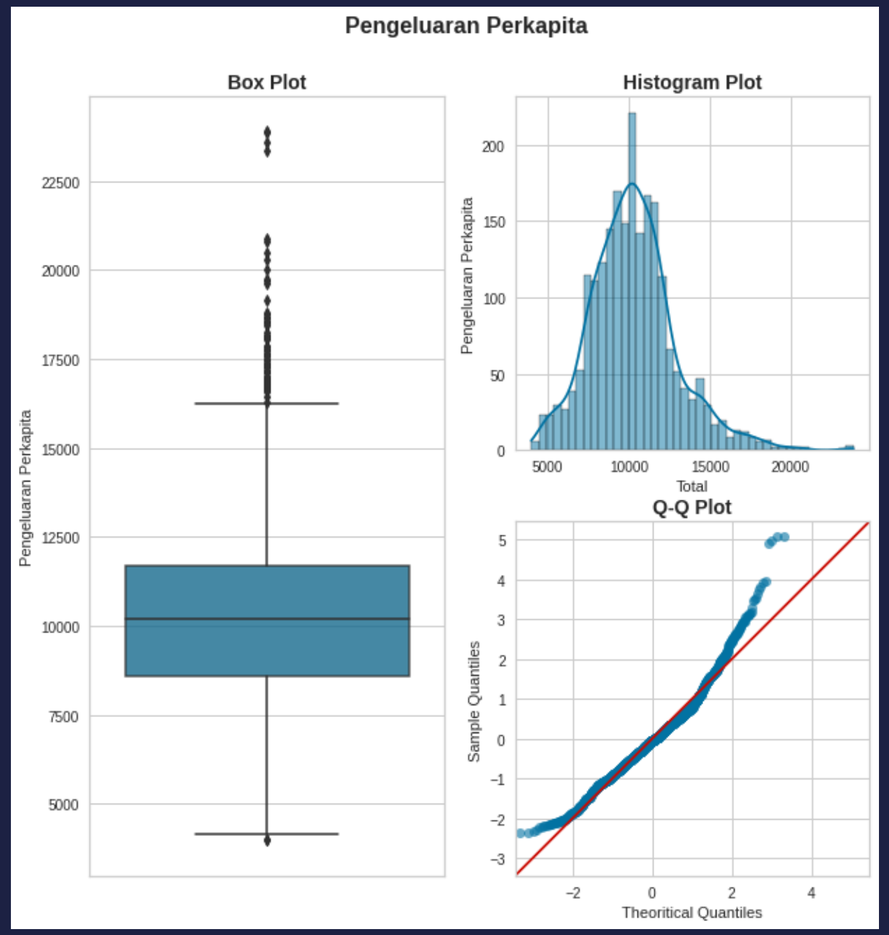
Terdapat outliers baik dibagian atas maupun bawah boxplot.

Berdasarkan histogram, variabel ini termasuk highly left skewed

Nilai kurtosis kolom ini 10.309. Artinya kolom ini termasuk leptokurtic.

Berdasarkan QQ-Plot, data menjauh dari 45 derajat di bagian bawah. Artinya data cenderung highly left skewed.

DISTRIBUSI DATA



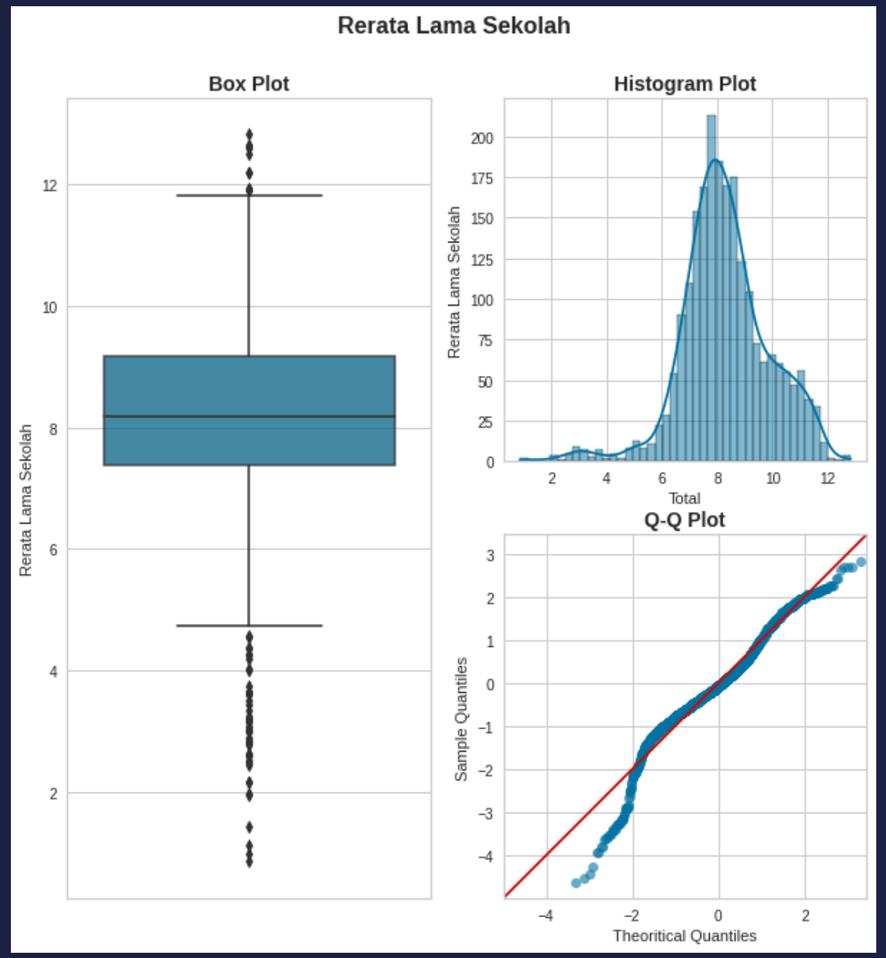
Berdasarkan boxplot, dataset memiliki outliers.

Berdasarkan histogram, kolom ini termasuk moderately right skewed.

Nilai kurtosis kolom ini 1.773. Artinya kolom ini termasuk platikurtic.

Berdasarkan QQ-Plot, data cenderung menjauh dari 45 derajat di bagian atas. Artinya data cenderung moderately right skewed.

DISTRIBUSI DATA

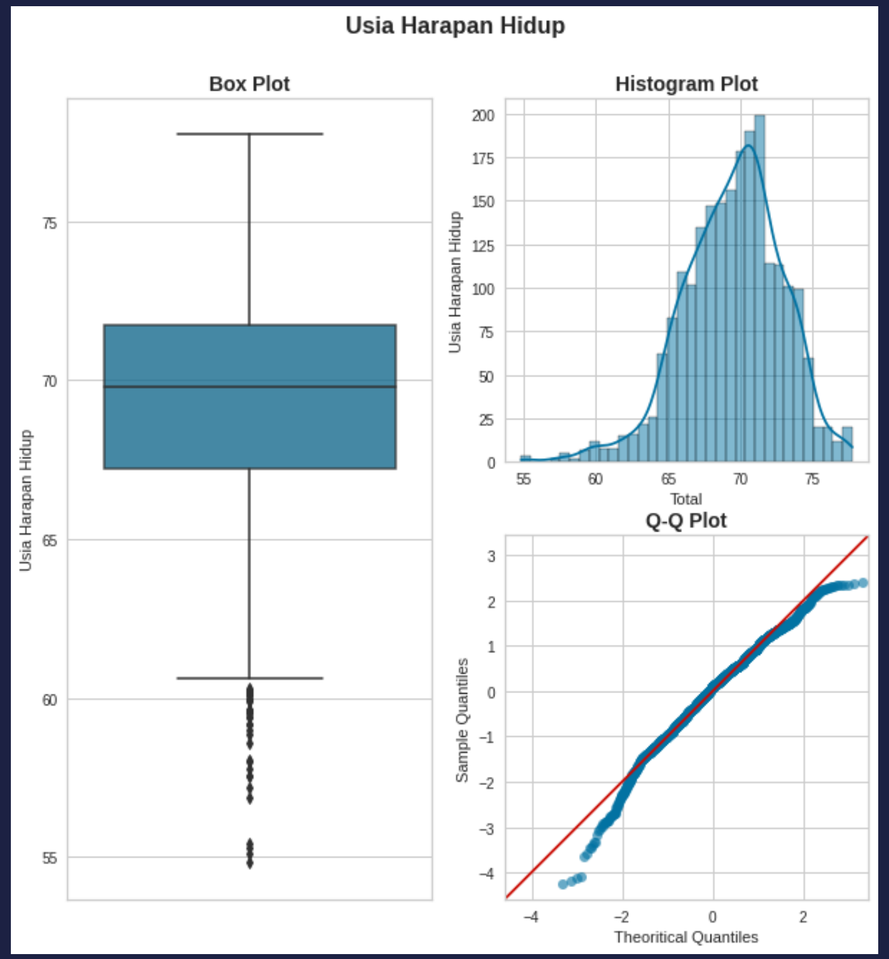


Berdasarkan boxplot, dataset memiliki outliers.

Berdasarkan histogram, kolom ini termasuk approximately symmetric.

Nilai kurtosis kolom ini 1.729. Artinya kolom ini termasuk platikurtic.

DISTRIBUSI DATA

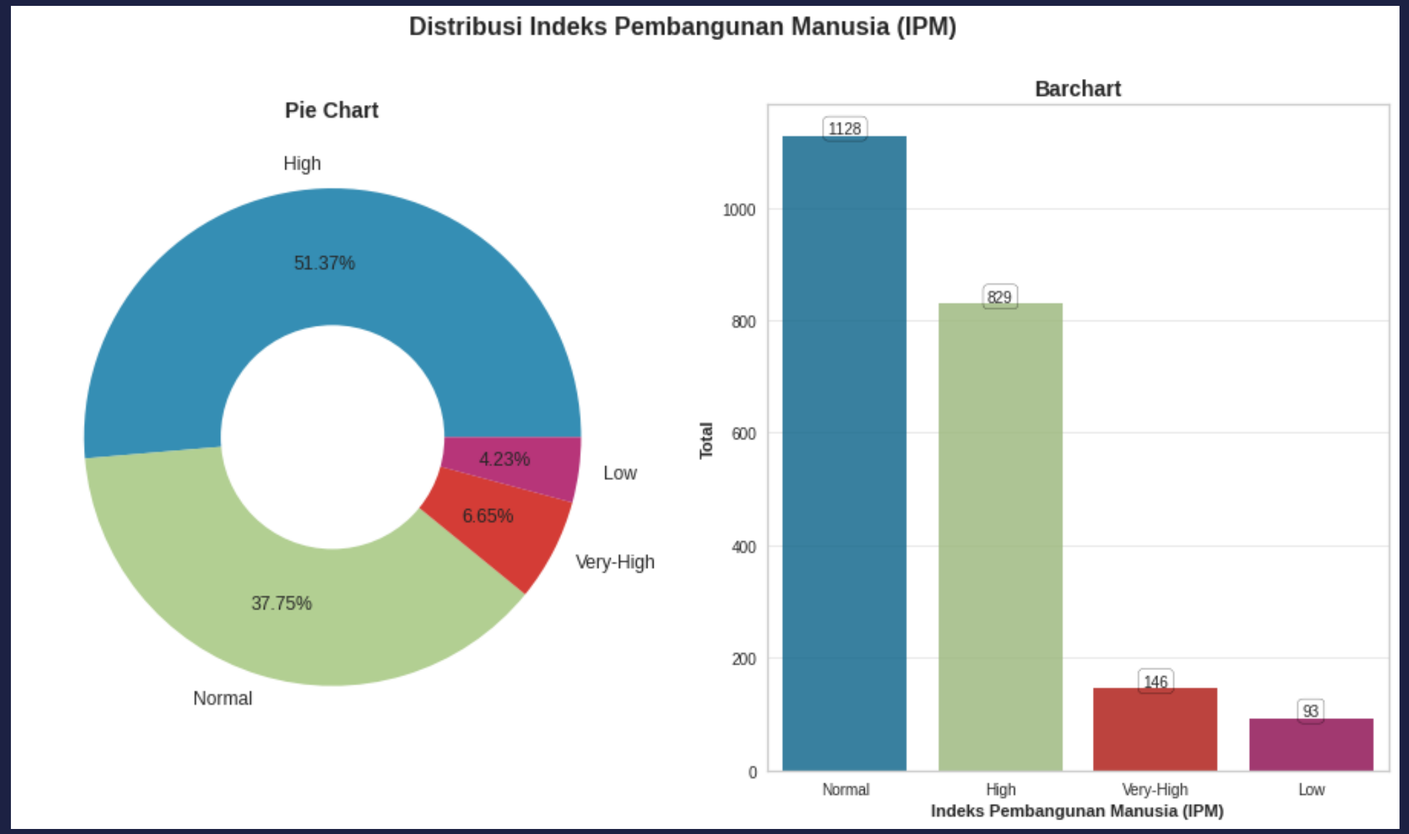


Berdasarkan boxplot, dataset memiliki outliers di bagian bawah boxplot.

Berdasarkan histogram, kolom ini termasuk approximately symmetric

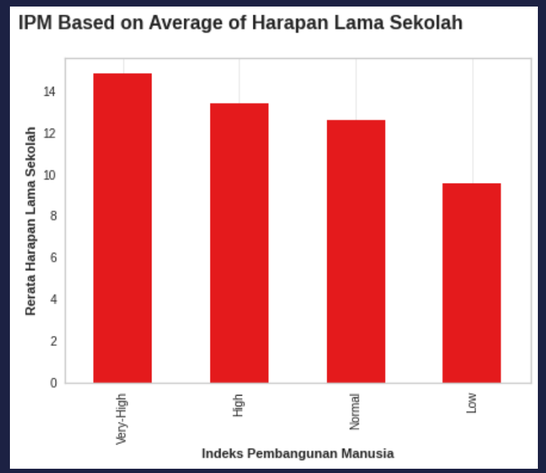
Nilai kurtosis kolom ini 0.686. Artinya kolom ini termasuk platikurtic.

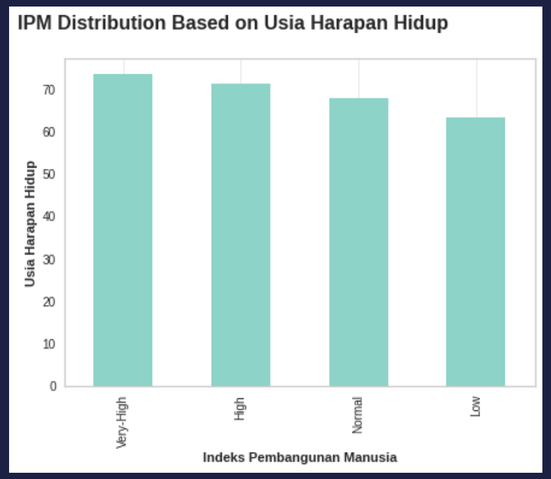
DISTRIBUSI DATA

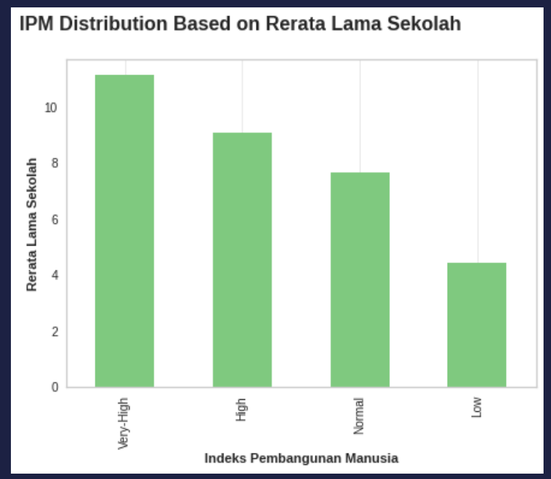


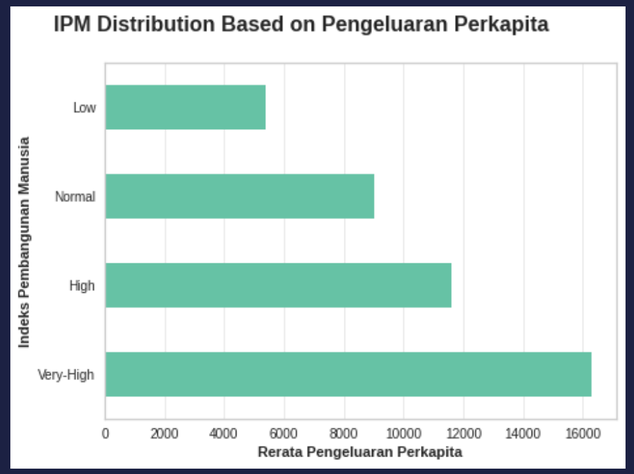
Jumlah data dengan IPM paling sedikit adalah Very-High dan Low dari total keseluruhan data.
Jumlah data IPM normal terbanyak di antara lainnya.

EXPLORATORY DATA









Adanya korelasi yang cukup signifikan antara rerata lama sekolah dengan IPM. Serta, pengeluaran perkapitan dengan IPM

PREPROCESING DATA

KONVERSI DATA TARGET MENJADI NUMERIK

```
df['IPM'].unique()
```

```
array(['High', 'Normal', 'Very-High', 'Low'], dtype=object)
```

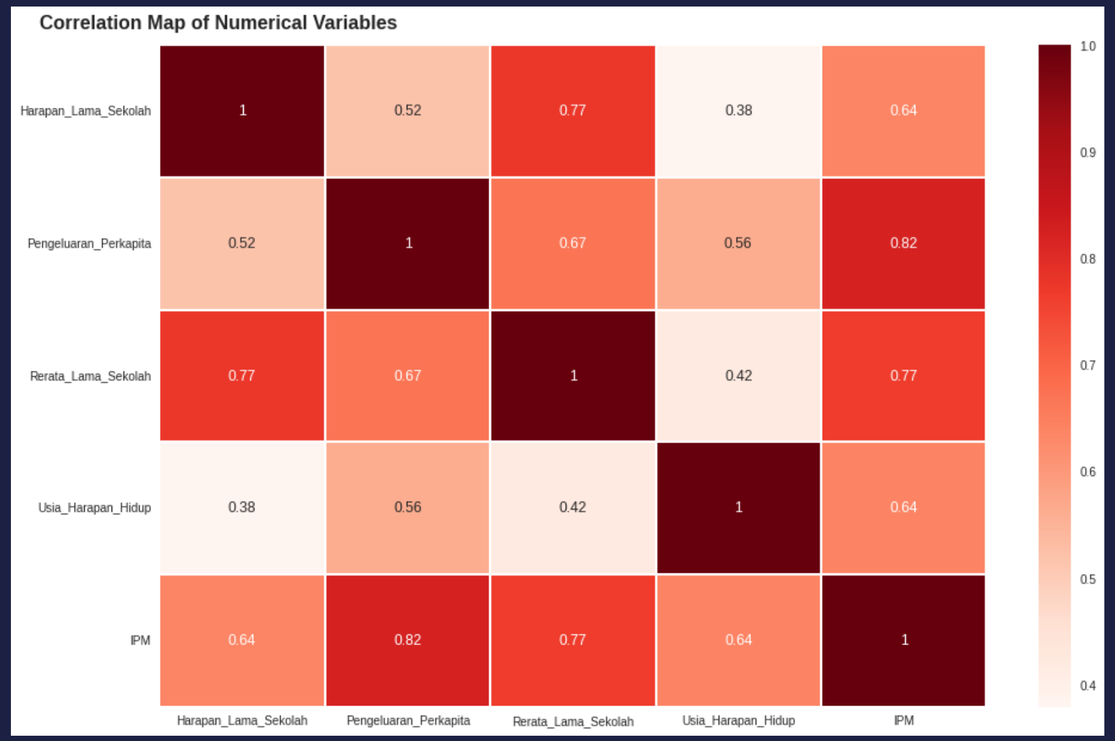
```
df['IPM'].replace(['Low', 'Normal', 'High', 'Very-High'],
                  [0, 1, 2, 3], inplace=True)
```

```
df['IPM'].unique()
```

```
array([2, 1, 3, 0])
```

PREPROCESING DATA

HEATMAP CORRELATION



Berdasarkan heatmap, diperoleh variabel `Pengeluaran_Perkapita` dan `Rerata_Lama_Sekolah` cukup berkorelasi secara positif dengan IPM.

PREPROCESING DATA

CEK DUPLIKAT VALUE & SPLIT DATA X DAN Y

```
print("Jumlah data yang duplikat: ", str(df.duplicated().sum()))
```

```
Jumlah data yang duplikat:  0
```

```
X = df.drop(['IPM'], axis=1)
y = df['IPM']
```

PREPROCESING DATA

STANDARISASI & SPLIT TRAIN DAN TEST SET

```
scaler = StandardScaler()
scaled_X = scaler.fit_transform(X)
```

```
scaled_X
```

```
array([[ 1.08824282, -0.28194717,  0.66945354,  0.1433277 ],
       [ 0.73781156, -1.19181217,  0.73773882, -1.21842258],
       [ 1.05777054, -0.58073122,  0.24111861, -0.59283217],
       ...,
       [-2.3932156 , -2.08778895, -3.29109263, -1.34936011],
       [ 1.56818129,  1.72621194,  1.8675498 ,  0.19861243],
       [-0.01637746,  0.27620845, -0.07547677,  0.50413333]])
```

```
X_test, X_train, y_test, y_train = train_test_split(scaled_X, y, stratify=y, test_size=0.2, random_state=42)
```

DECISION TREE

criterion='gini', max_depth=8,
random_state=42

RANDOM FOREST

criterion='gini', max_depth=10,
random_state=42

ADABOOST

dt =
DecisionTreeClassifier(criterion='gini',
max_depth=8, random_state=42)

dt, n_estimators=100,
random_state=42

XGBOOST

Default

STACKING

```
estimators = [
('clf1' , KNeighborsClassifier()),
('clf2' , GaussianNB()),
('clf3' , SVC())]


estimators=estimators,
final_estimator=LogisticRegression(random_state=42)
```

EVALUATION

93%

ADABOOST

95.6%

RANDOM FOREST

97.8%

STACKING

KNN, Naive Bayes, SVM sebagai estimator dan Logistic Regression sebagai final estimator

95.1%

XGBOOST

92.5%

DECISION TREE

Kesimpulan

1. Persebaran data dari setiap variabel cenderung tidak terdistribusi normal.
2. Variabel `Pengeluaran_Perkapita` dan `Rerata_Lama_Sekolah` cukup berkorelasi secara positif dengan IPM.
3. Model dengan teknik stacking merupakan model dengan akurasi paling tinggi sebesar 97.8%.

TERIMA KASIH

Fayza Apriliza - Broyden