

# **Analisa Studi Kasus Tingkat IPM menggunakan Metode Klasifikasi**

Ditulis oleh Giselle Halim (Kelas Tensor)

# Latar Belakang

Indeks Pembangunan Manusia (IPM) adalah indeks yang menjelaskan bagaimana penduduk dapat mengakses hasil pembangunan dalam memperoleh pendapatan, kesehatan, pendidikan, dan sebagainya. IPM diperkenalkan oleh United Nations Development Programme (UNDP) pada tahun 1990 dan dipublikasikan secara berkala dalam laporan tahunan Human Development Report (HDR). IPM merupakan indikator penting untuk mengukur keberhasilan dalam upaya membangun kualitas hidup manusia (masyarakat/penduduk) dan dapat menentukan peringkat atau level pembangunan suatu wilayah/negara.

Penelitian ini bertujuan untuk melihat tingkat IPM yang pada data dikategorikan sebagai Low, Normal, High, dan Very High dengan menggunakan metode klasifikasi agar dapat dilakukan eksplorasi data dan memprediksi tingkat IPM pada data testing. Selain itu, penelitian ini akan membandingkan tiga model algoritma, yaitu Random Forest, Gradient Boost, dan XGboost untuk melihat hasil dan akurasi dari ketiga model tersebut.

# Rumusan Masalah



## Rumusan Masalah 1

Bagaimana hasil eksplorasi pada data?

## Rumusan Masalah 2

Bagaimana cara memprediksi tingkat IPM pada data dengan metode klasifikasi (ensemble learning)?

## Rumusan Masalah 3

Bagaimana hasil perbandingan algoritma yang digunakan berdasarkan hasil dan akurasi masing-masing algoritma?

# Data yang digunakan

Data yang digunakan adalah data mengenai Indeks Pembangunan Manusia. Data ini terdiri dari 2196 baris data dan telah terbebas dari data duplikat maupun missing values. Tipe data juga telah sesuai dengan kebutuhan.

Variabel terikat atau dependen pada data adalah tingkat IPM yang terbagi menjadi 4 kategori yang pada fase training akan dikonversikan menjadi label numerik, kategori IPM pada data adalah:

- Low (rendah), dilambangkan dengan angka 1 pada data
- Normal (sedang), dilambangkan dengan angka 2 pada data
- High (tinggi), dilambangkan dengan angka 0 pada data
- Very High (sangat tinggi), dilambangkan dengan angka 3 pada data

Variabel bebas atau independen pada data terdiri dari: Harapan lama sekolah, rerata lama sekolah, pengeluaran perkapita, dan usia harapan hidup.

# Data Preprocessing

## Data Cleaning

Pada saat data cleaning, tidak ada data duplikat atau kosong, tipe data juga sudah sesuai dengan yang dibutuhkan sehingga data bisa dikatakan bersih.

## Data Exploration

Pada hasil eksplorasi, ditemukan beberapa insight yang akan dijelaskan lebih lanjut.

## Data Training

Pada data training, data dibagi menjadi data training dan data testing dengan rasio 60:40. Setelahnya, dilakukan feature scaling dan balancing data. Untuk membuat model, digunakan algoritma random forest, gradient boosting, dan XGboost.



# Data Exploration

## Harapan Lama Sekolah

Harapan lama sekolah pada data memiliki rata-rata 12,9 tahun dengan waktu tersingkat 0,85 tahun dan waktu terpanjang 12,83 tahun.

## Pengeluaran per Kapita

Pengeluaran per kapita pada data memiliki rata-rata 10.323 dengan pengeluaran paling sedikit sebesar 3.975 dan pengeluaran tertinggi sebesar 23.888.

## Rerata Lama Sekolah

Rerata lama sekolah pada data memiliki rata-rata 8,2 tahun dengan waktu tersingkat 2,95 tahun dan waktu terpanjang 17,8 tahun.

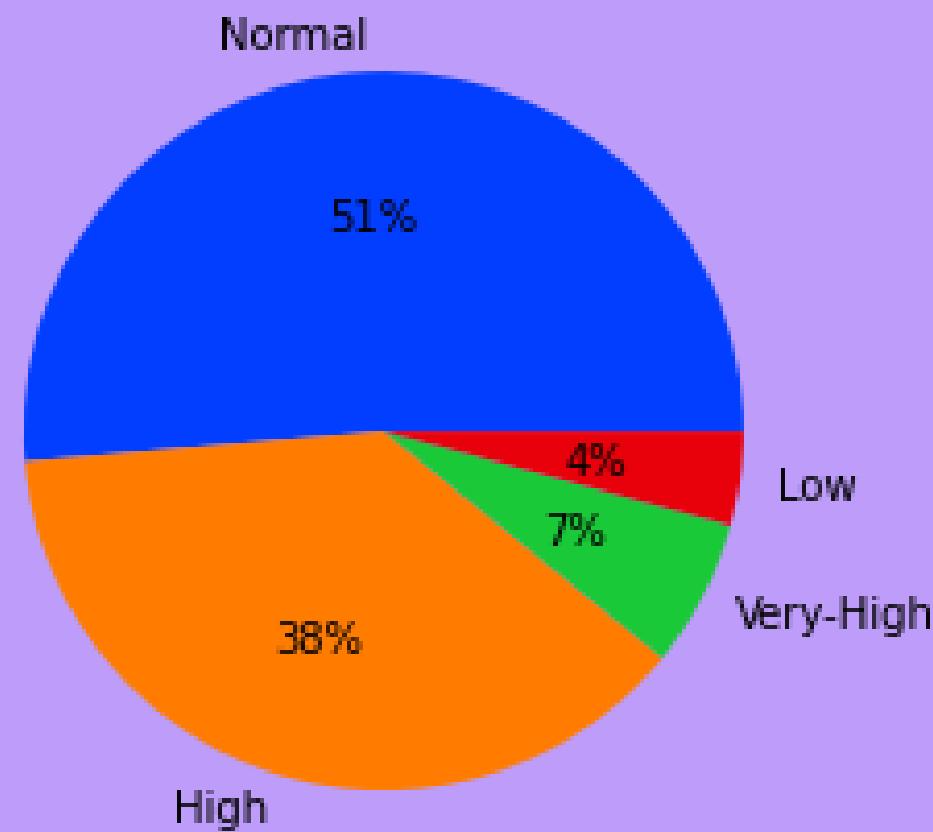
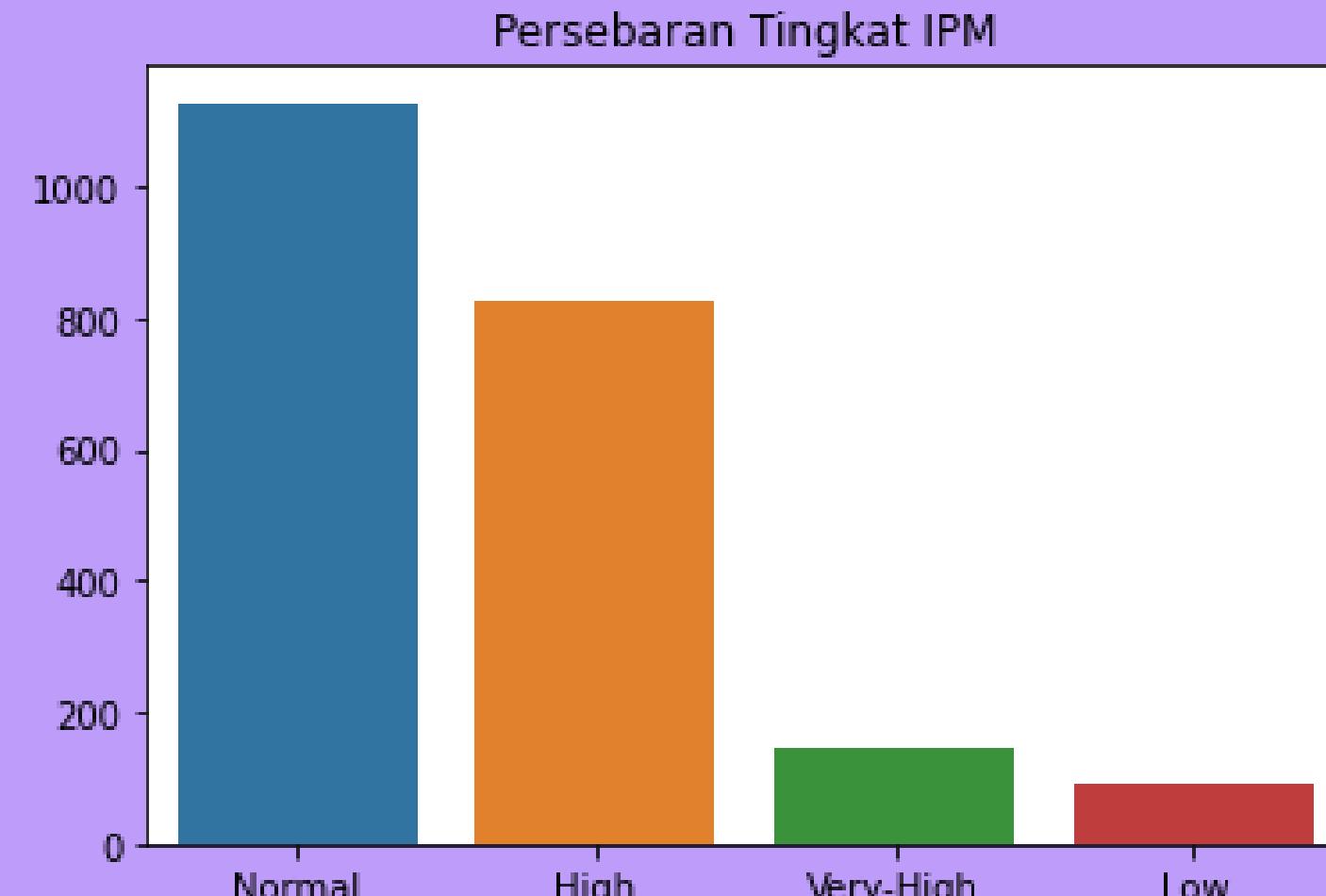
## Usia Harapan Hidup

Usia harapan hidup pada data memiliki rata-rata 69,5 tahun dengan usia tersingkat 54,82 tahun dan usia terlama 77,73 tahun.

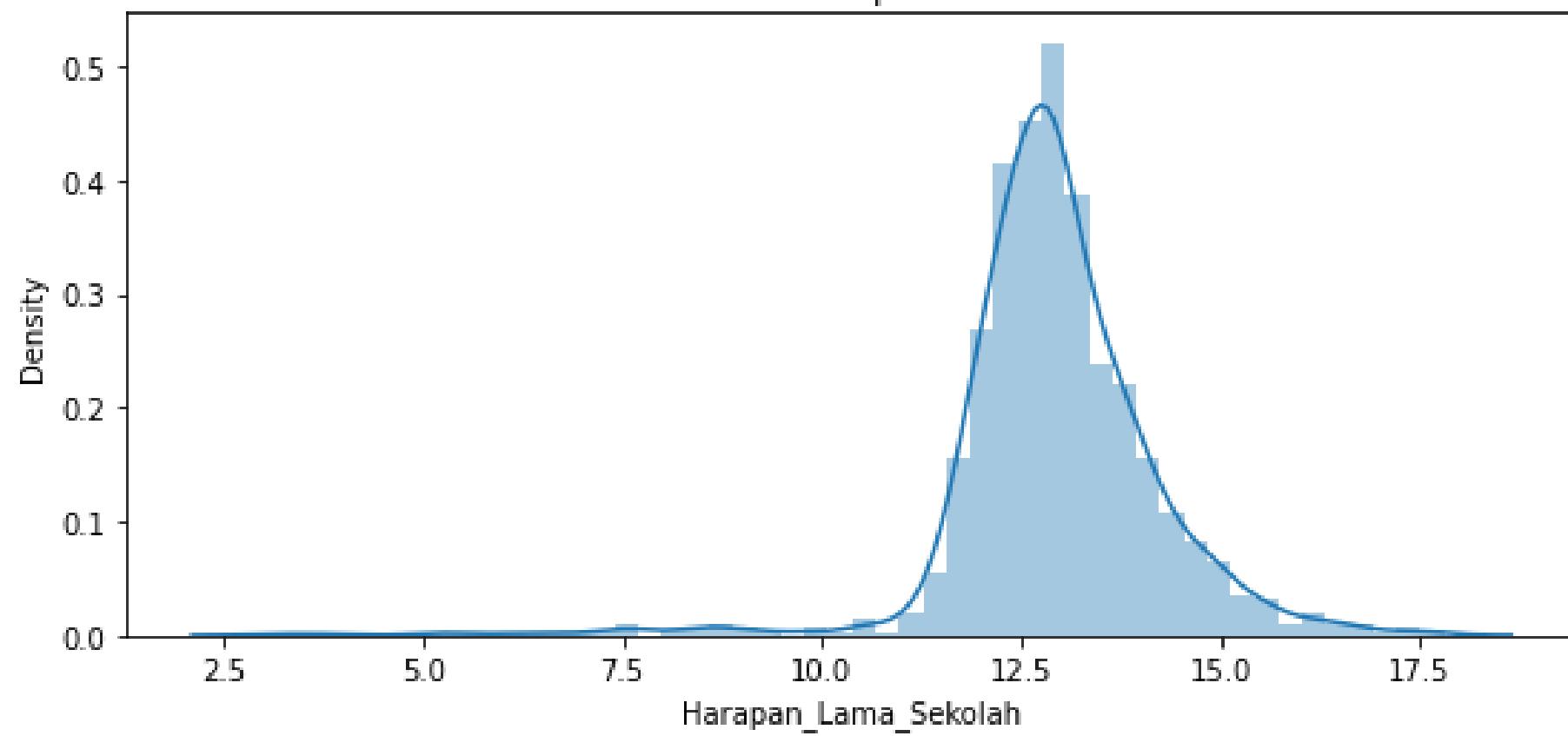
# Grafik Data

Pada data klasifikasi IPM, terlihat bahwa data setiap kategori memiliki jarak yang cukup jauh, terutama data low dan very high. Jumlah persebaran data adalah sebagai berikut:

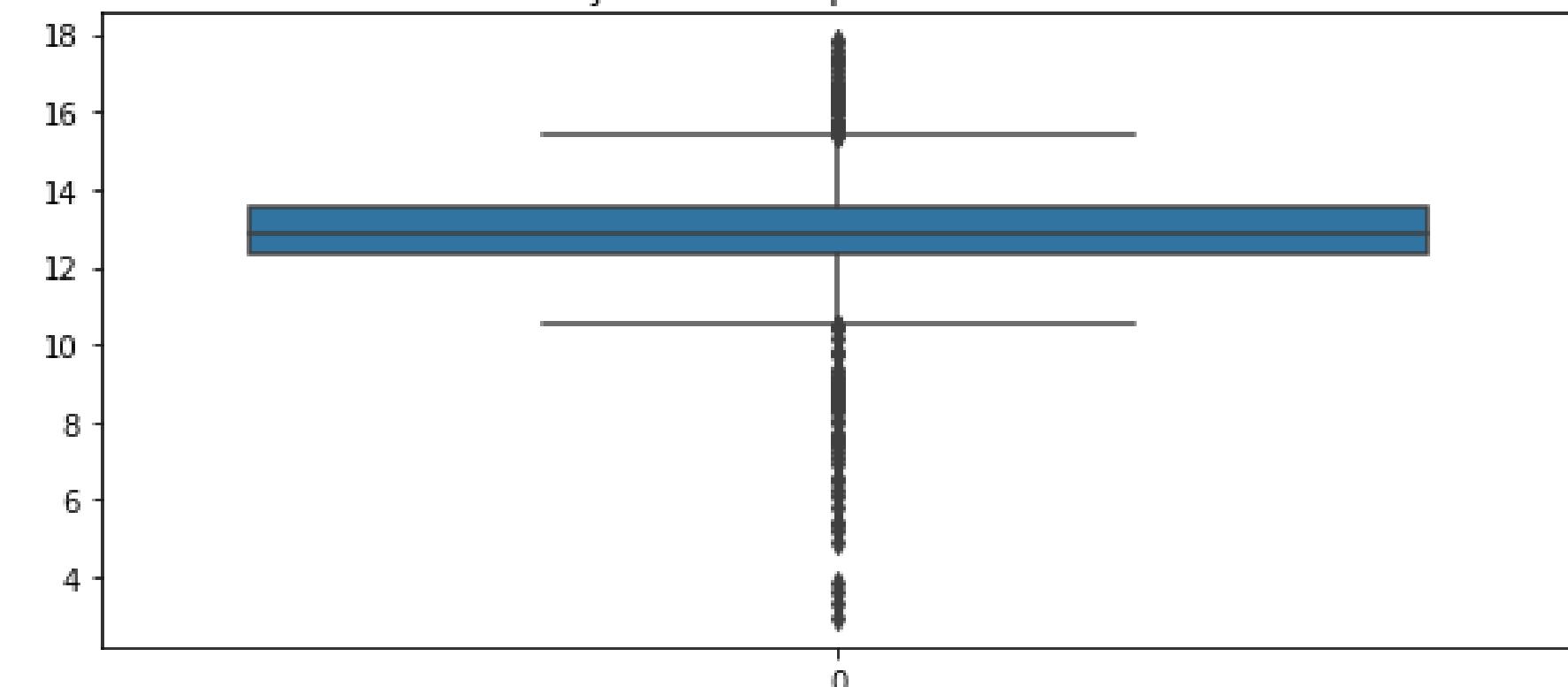
- Low = 93 (4%)
- Normal = 1128 (51%)
- High = 829 (38%)
- Very High = 146 (7%)



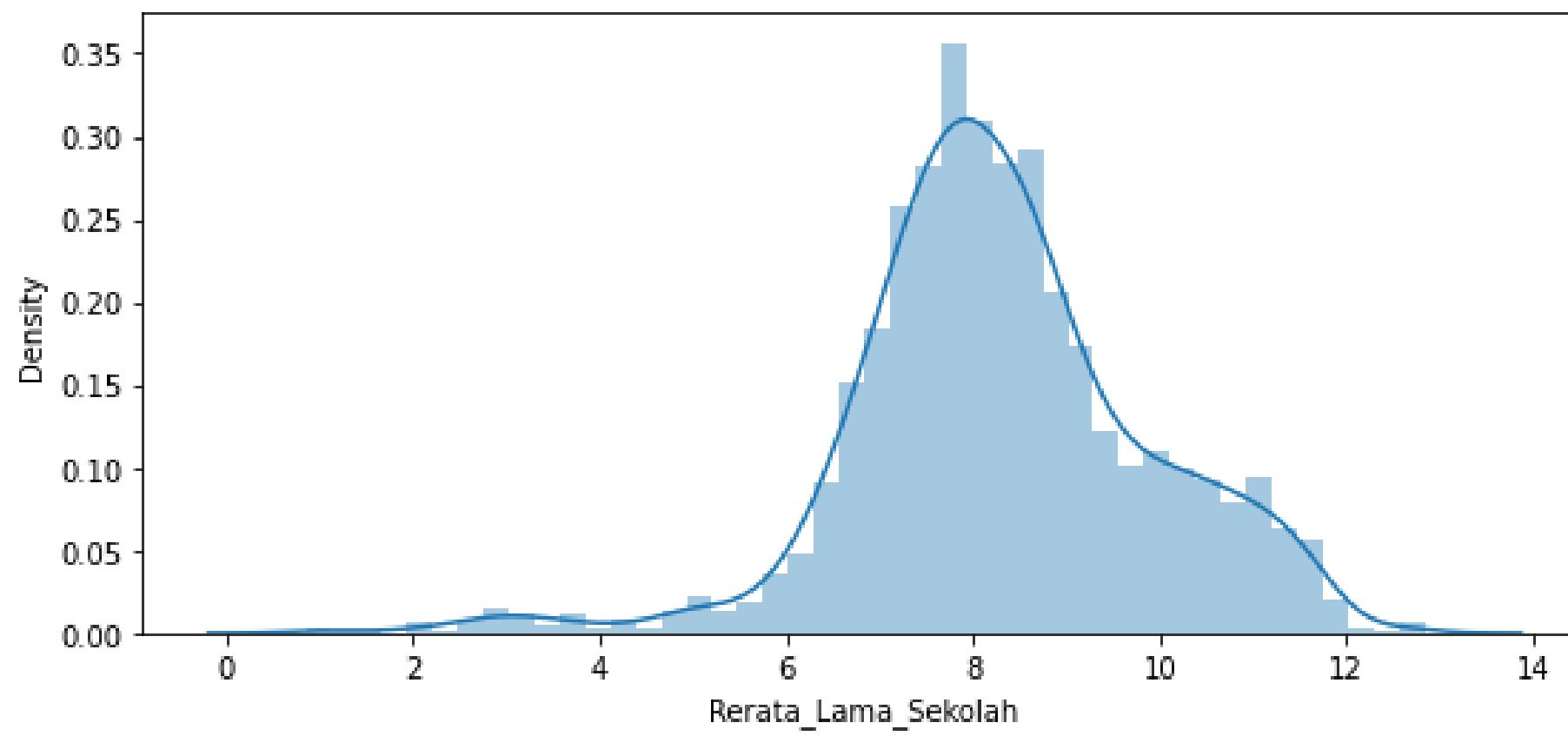
Tabel Distribusi Harapan Lama Sekolah



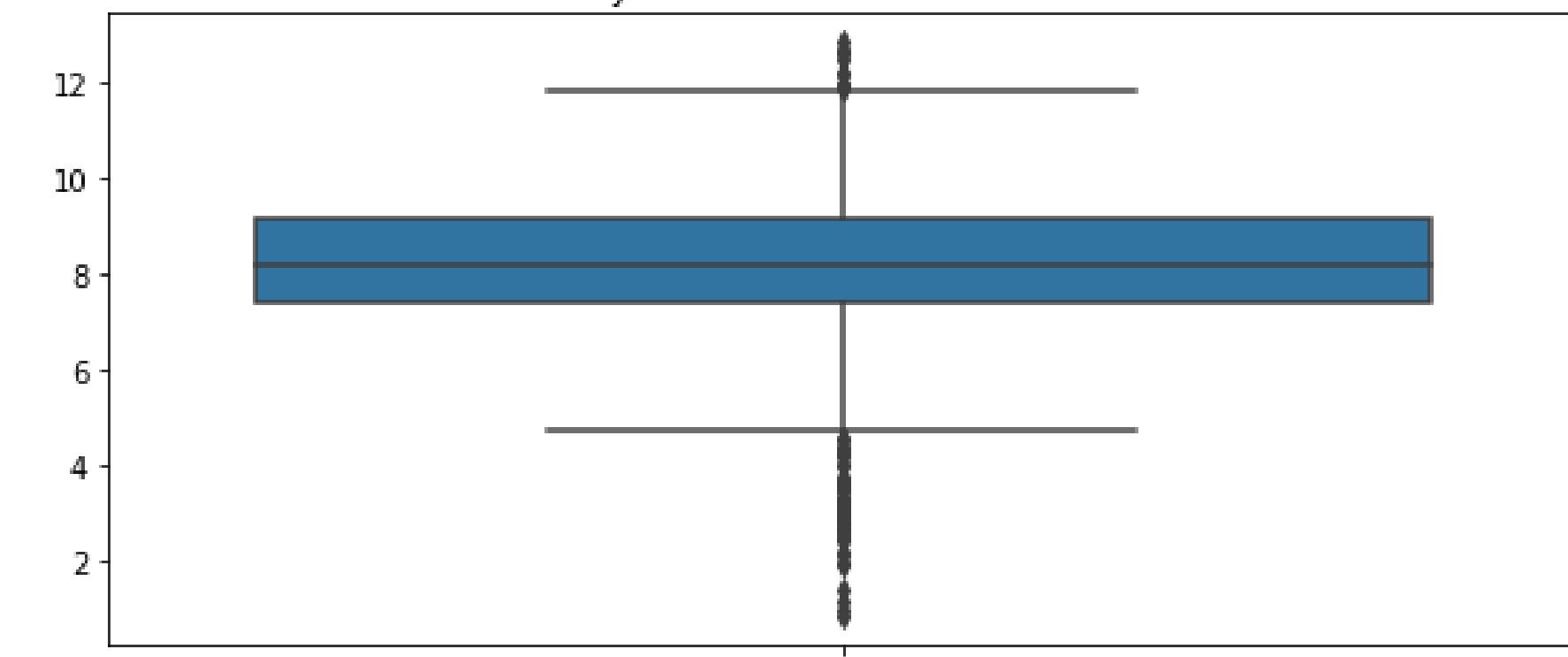
Penyebaran Harapan Lama Sekolah



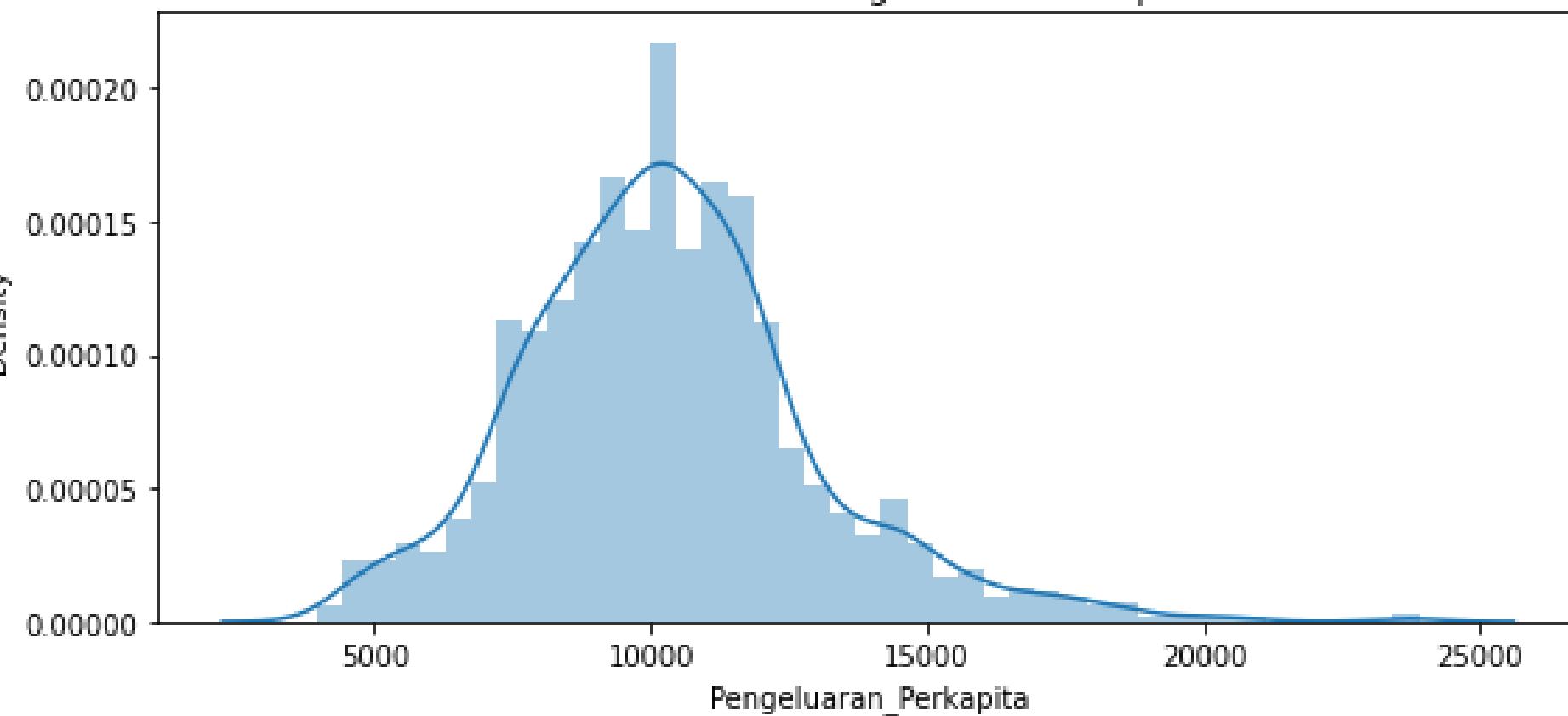
Tabel Distribusi Rerata Lama Sekolah



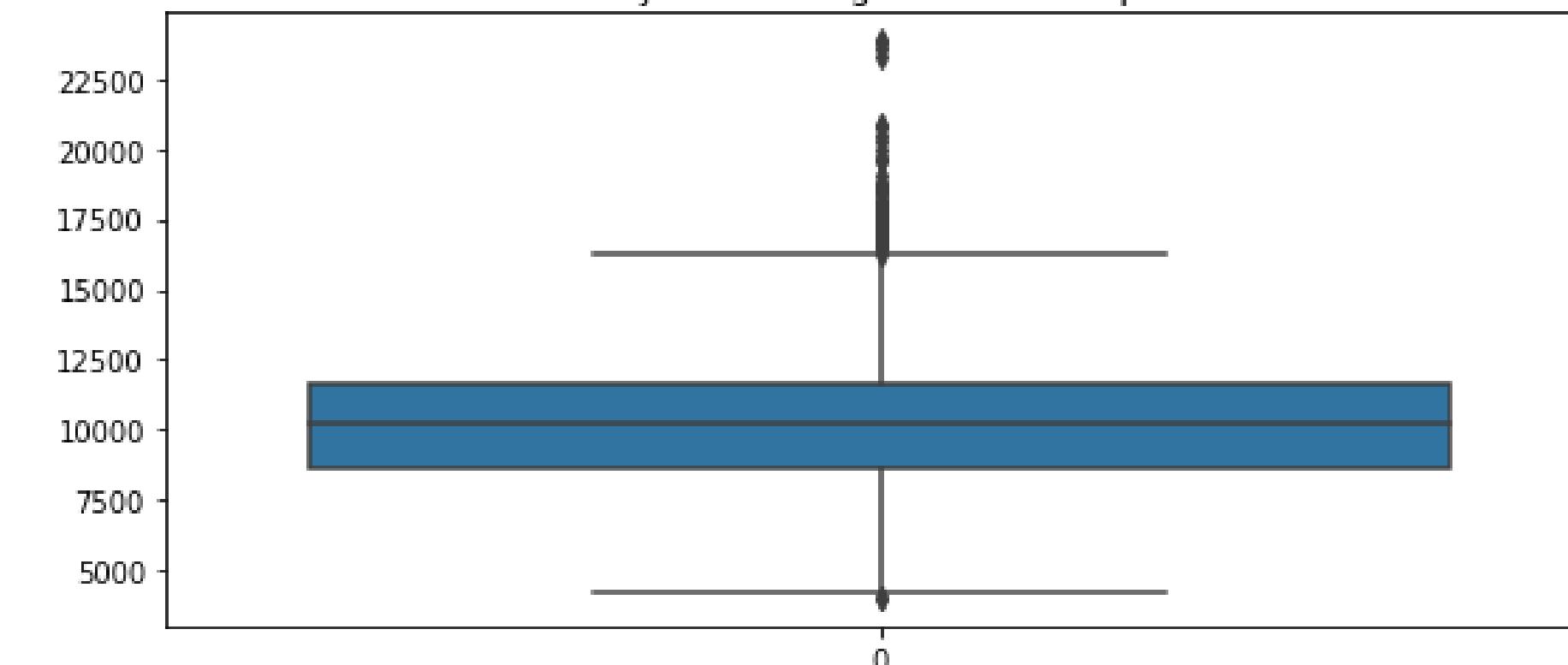
Penyebaran Rerata Lama Sekolah



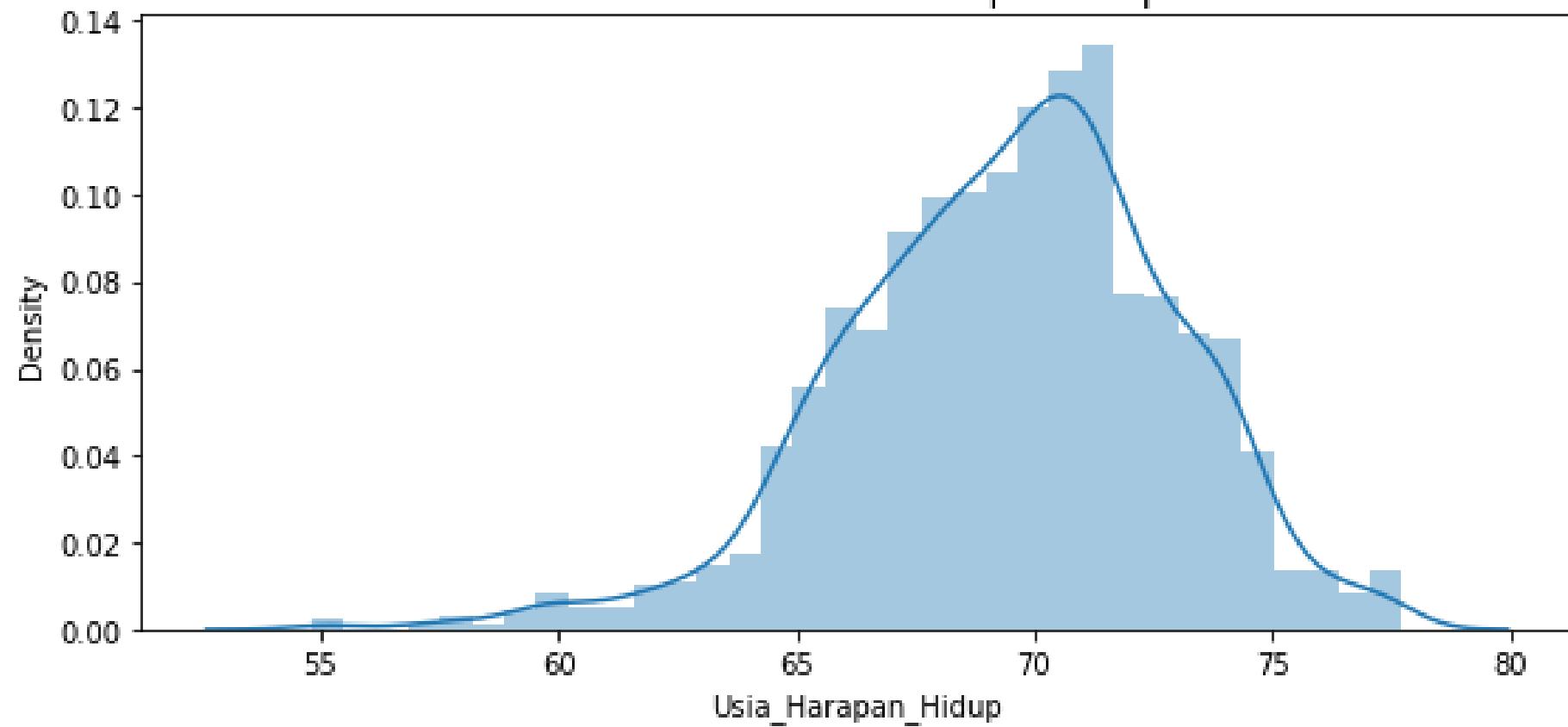
Tabel Distribusi Pengeluaran Per Kapita



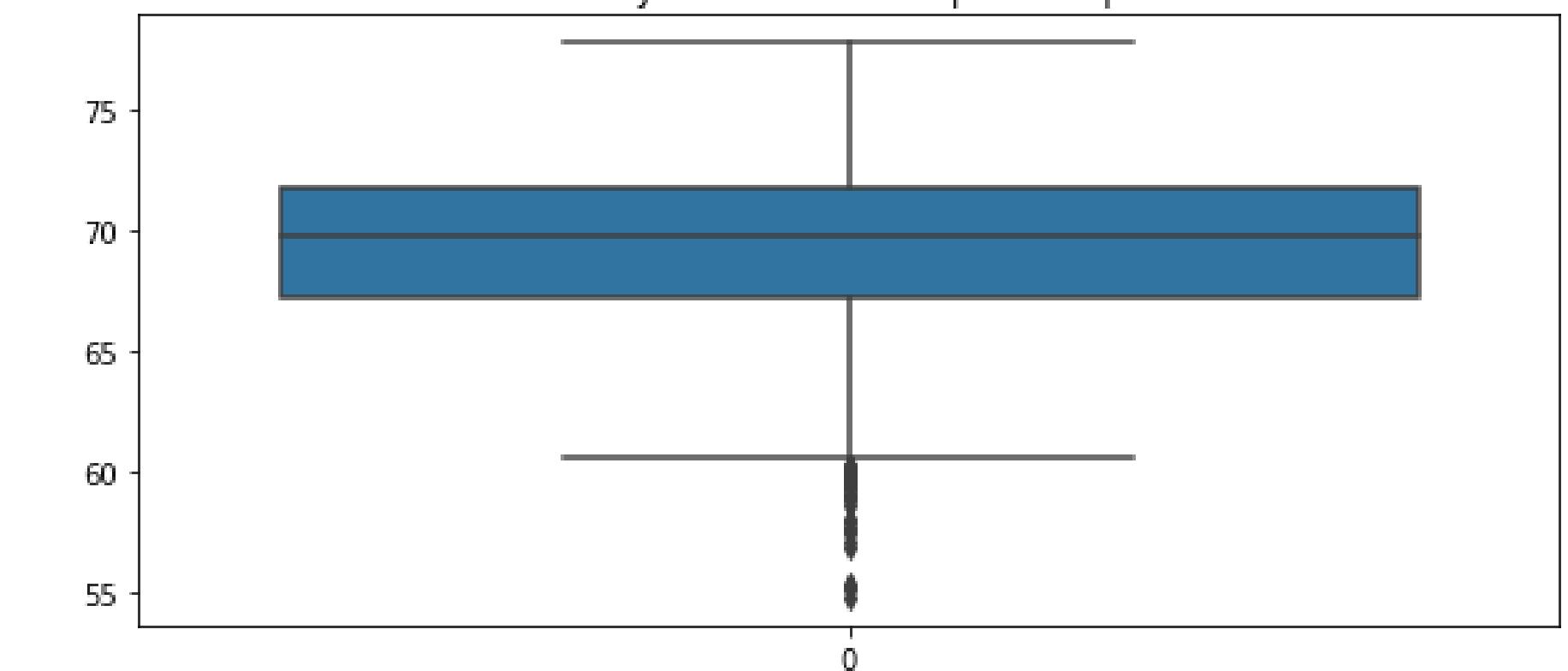
Penyebaran Pengeluaran Per Kapita



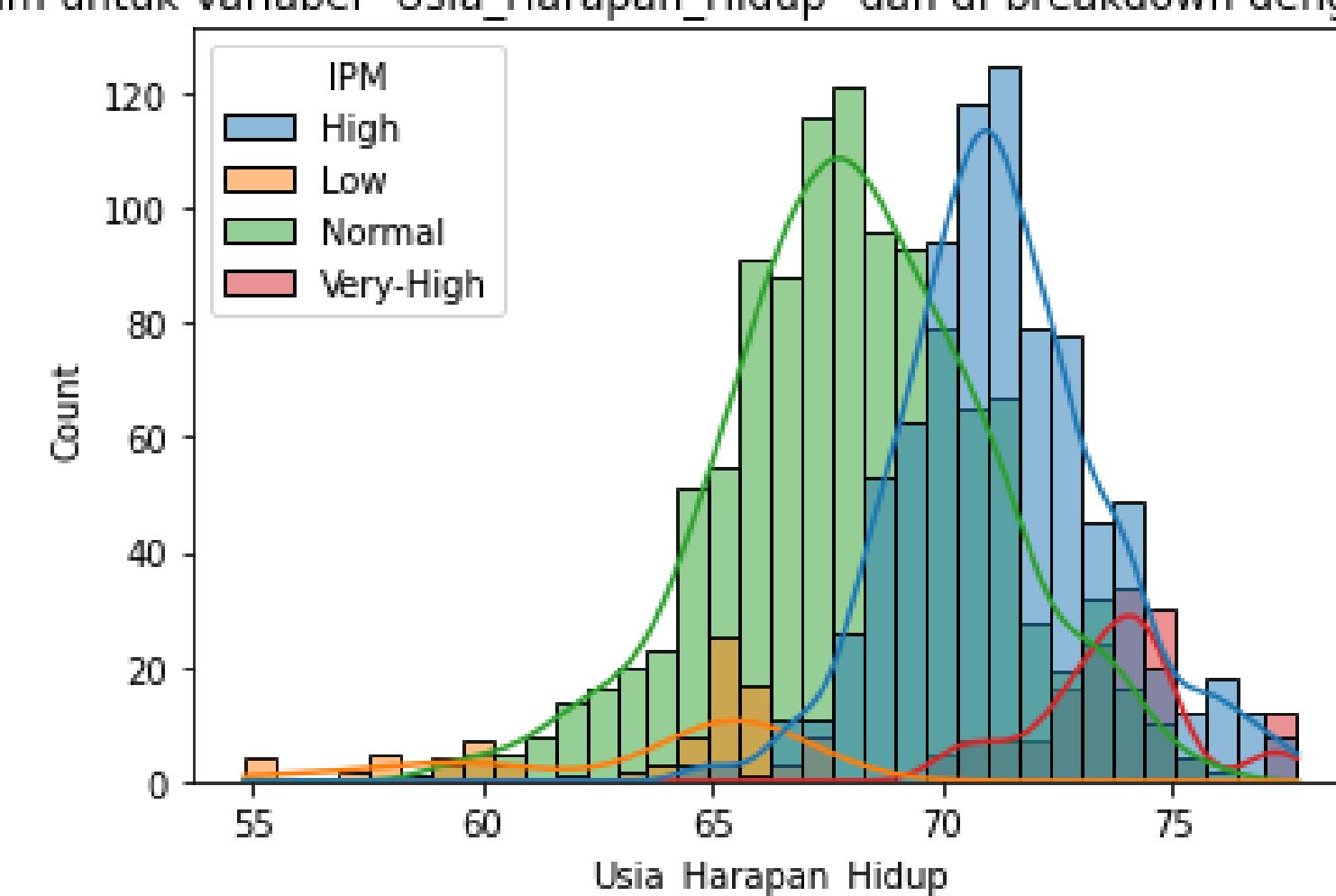
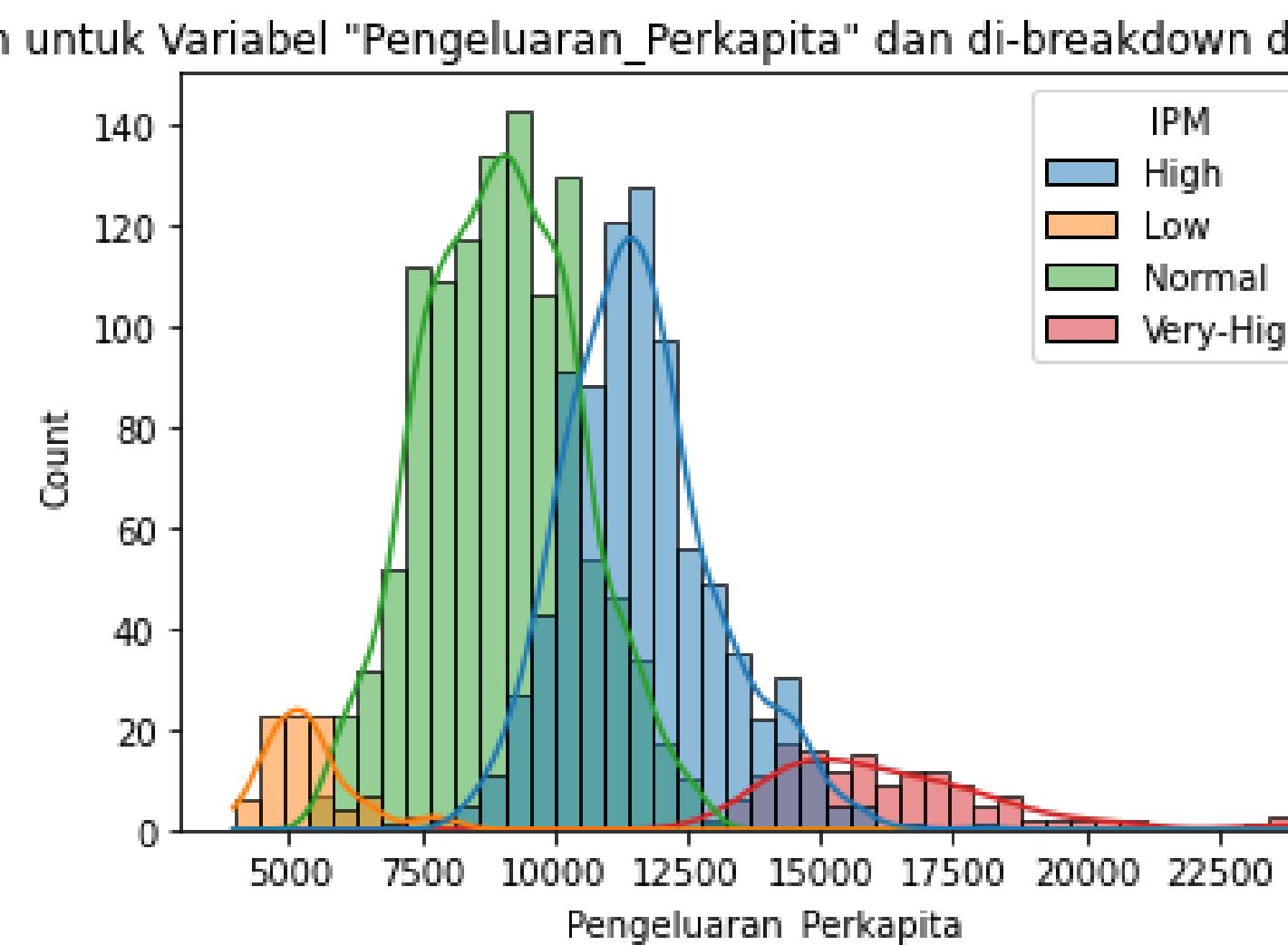
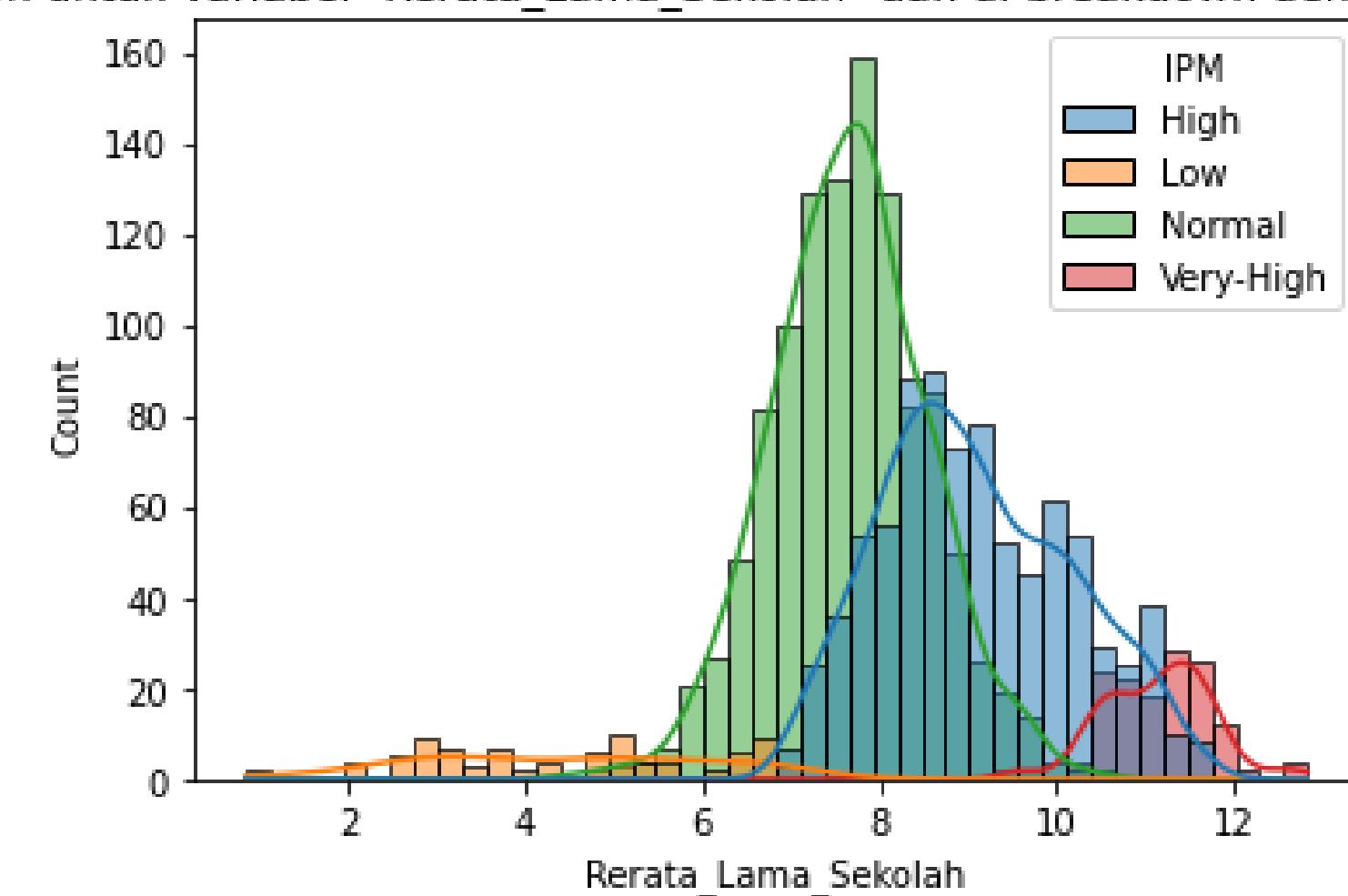
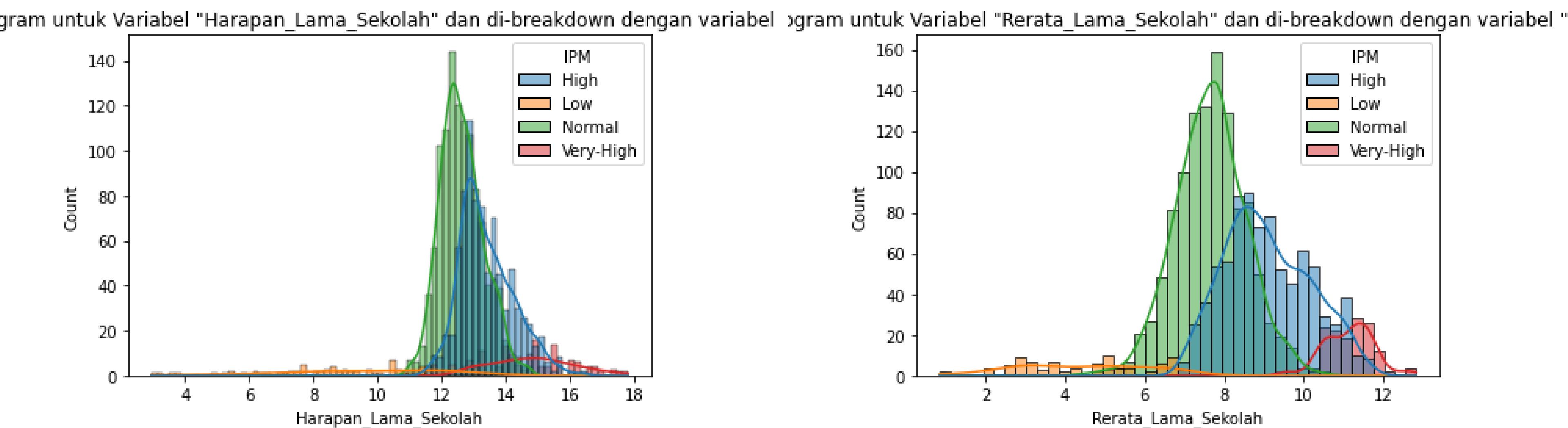
Tabel Distribusi Usia Harapan Hidup



Penyebaran Usia Harapan Hidup









- Dari 2196 data, ada 1908 data yang memiliki harapan lama sekolah diatas 12 tahun, tetapi hanya ada 7 data yang mencapai rerata lama sekolah selama itu. Sekitar 2189 data memiliki rerata lama sekolah dibawah 12 tahun.
- Usia harapan hidup rata-rata menurut Biro Pusat Statistik (BPS) pada tahun 2021 adalah 73,5 tahun. Ada 261 data yang diatas rata-rata dan 1935 data memiliki usia harapan hidup dibawah itu.
- Hasil perhitungan korelasi antar variabel adalah sebagai berikut:

	Harapan_Lama_Sekolah	Pengeluaran_Perkapita	Rerata_Lama_Sekolah	Usia_Harapan_Hidup
Harapan_Lama_Sekolah	1.00	0.52	0.77	0.38
Pengeluaran_Perkapita	0.52	1.00	0.67	0.56
Rerata_Lama_Sekolah	0.77	0.67	1.00	0.42
Usia_Harapan_Hidup	0.38	0.56	0.42	1.00

- Terlihat bahwa harapan lama sekolah memiliki korelasi yang kuat dengan rerata lama sekolah dengan nilai 0,77.
- Pengeluaran per kapita dan rerata lama sekolah memiliki korelasi yang cenderung kuat dengan nilai 0,67.
- Usia harapan hidup dan pengeluaran per kapita memiliki korelasi yang cukup kuat dengan nilai 0,56.

# Model yang digunakan

## **Random Forest**

Digunakan random forest dengan jumlah tree (n\_estimators) sebanyak 35 dan kriteria entropy.

## **XGboost**

Digunakan XGboost dengan jumlah n\_estimators sebanyak 35.

## **Gradient Boosting**

Digunakan gradient boosting dengan jumlah n\_estimators sebanyak 35.

**Agar data balance, dilakukan balancing data dengan SMOTE. Evaluasi data dilakukan dengan perhitungan akurasi model training dan testing, confusion matrix, dan classification report.**

```
from sklearn.ensemble import RandomForestClassifier

classifier_rf = RandomForestClassifier(n_estimators=35, criterion="entropy")
classifier_rf.fit(X_train, y_train)
y_pred_rf = classifier_rf.predict(X_test)
```

```
from sklearn.ensemble import GradientBoostingClassifier

classifier_gb = GradientBoostingClassifier(n_estimators=35)
classifier_gb.fit(X_train, y_train)
y_pred_gb = classifier_gb.predict(X_test)
```

```
from xgboost import XGBClassifier

classifier_xgb = XGBClassifier(n_estimators=35)
classifier_xgb.fit(X_train, y_train)
y_pred_xgb = classifier_xgb.predict(X_test)
```

```
from imblearn.over_sampling import SMOTE
# define oversampling strategy
SMOTE = SMOTE()

# fit and apply the transform
X_train_SMOTE, y_train_SMOTE = SMOTE.fit_resample(X_train, y_train)
```

# Perbandingan Model

Hasil perbandingan dari pengujian ketiga model **sebelum** dilakukan balancing data dengan SMOTE adalah sebagai berikut:

- Random Forest memiliki akurasi training sebesar 100% dan akurasi testing sebesar 96,3%.
- Gradient Boosting memiliki akurasi training sebesar 99,7% dan akurasi testing sebesar 95,5%.
- XGboost memiliki akurasi training sebesar 100% dan akurasi testing sebesar 96,9%.

Hasil perbandingan dari pengujian ketiga model **setelah** dilakukan balancing data dengan SMOTE adalah sebagai berikut:

- Random Forest memiliki akurasi training sebesar 100% dan akurasi testing sebesar 96,7%.
- Gradient Boosting memiliki akurasi training sebesar 99,6% dan akurasi testing sebesar 94,7%.
- XGboost memiliki akurasi training sebesar 100% dan akurasi testing sebesar 96,5%.

# Perbandingan Model

Jika dilihat dengan nilai classification report dan akurasi saat pengujian maka setelah dilakukan balancing data, model Random Forest memiliki akurasi testing dan nilai classification report terbaik dibandingkan 2 model lainnya. Model terbaik kedua adalah XGboost, lalu yang terakhir adalah Gradient Boosting.

Tetapi sebenarnya nilai akurasi ketiga model tidak berbeda jauh, model terbaik hanya 2% lebih akurat dibandingkan model yang berada diurutan ketiga.

Selain itu, juga dapat terlihat bahwa balancing data dapat menurunkan akurasi training dan testing seperti pada terlihat pada model Gradient Boosting dan XGboost.

```
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred_rf))
```

	precision	recall	f1-score	support
0	0.97	0.96	0.96	340
1	0.97	0.97	0.97	37
2	0.98	0.98	0.98	440
3	0.87	1.00	0.93	62
accuracy			0.97	879
macro avg	0.95	0.98	0.96	879
weighted avg	0.97	0.97	0.97	879

```
#Classification report
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred_xgb))
```

	precision	recall	f1-score	support
0	0.97	0.94	0.95	340
1	0.97	0.97	0.97	37
2	0.98	0.97	0.98	440
3	0.82	1.00	0.90	62
accuracy			0.96	879
macro avg	0.94	0.97	0.95	879
weighted avg	0.96	0.96	0.96	879

```
#Classification report
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred_gb))
```

	precision	recall	f1-score	support
0	0.96	0.93	0.94	340
1	0.95	0.97	0.96	37
2	0.97	0.97	0.97	440
3	0.83	1.00	0.91	62
accuracy			0.95	879
macro avg	0.93	0.97	0.94	879
weighted avg	0.96	0.95	0.95	879

# Kesimpulan

- Model Random Forest adalah model yang terbaik untuk data IPM yang digunakan.
- Tingkat IPM pada data adalah sebagai berikut: Low sebanyak 93 data (4%); Normal sebanyak 1128 data (51%); High sebanyak 829 data (38%); Very High sebanyak 146 data (7%).
- Semakin tinggi harapan lama sekolah, rerata lama sekolah, pengeluaran per kapita, dan usia harapan hidup, maka tingkat IPM cenderung semakin baik.
- Perbedaan data setiap kategori memiliki perbedaan yang cukup jauh dan data juga memiliki outlier.
- Masih banyak jumlah data yang rerata lama sekolahnya jauh dibawah harapan lama sekolah.
- Masih ada beberapa data yang usia harapan hidupnya dibawah rata-rata dari BPS.
- Terdapat korelasi yang kuat antara harapan lama sekolah dan rerata lama sekolah; pengeluaran per kapita dan rerata lama sekolah; serta usia harapan hidup dan pengeluaran per kapita.

# **THANK YOU**

[Google Colab Link](#)