CLASSIFICATION OF POVERTY LEVELS IN INDONESIA USING MACHINE LEARNING

Muhamad Habbiebie Robbi

January 2025

TABLE OF CONTENT

PROJECT DESCRIPTION

This project aims to build the best classification model for predicting poverty levels in Indonesia based on socio-economic and demographic features of each region. The dataset consists of 514 entries representing provinces and districts/cities in Indonesia with 12 features. These features include economic indicators, education levels, health metrics, and other key socio-economic statistics. The data is preprocessed and analyzed to ensure robust model development and evaluation.

TOOLS

PROJECT GOALS

1. Develop multiple classification models to predict poverty levels in Indonesia.
2. Evaluate and compare the models based on their accuracy to identify the best-performing model.
3. Analyze the relationships between socio-economic features and their impact on poverty prediction.

METHODOLOGY

01 Load Data

02 Exploratory Data Analysis (EDA)

03 Data Preprocessing

04 Split data

05 Modeling

06 Predict and Evaluate

METHODOLOGY

LOAD DATA

Import Library

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
```

```
file_path = "D:/Documents/UNJEHH/SEMESTER 6/BOOTCAMP/DIBIMBING/Data/Klasifikasi Tingkat Kemiskinan di Indor
data = pd.read_excel(file_path)
```

METHODOLOGY

EXPLORATORY DATA ANALYS

```
Missing Values:
 Provinsi                          0
Kab/Kota                           0
P0 Persen                          0
Mean Lama Sekolah 15+              0
Pengeluaran perKapita              0
IPM                                0
Umur Harapan Hidup                 0
Akses Sanitasi Layak               0
Akses Air Minum Layak              0
Pengangguran Terbuka               0
Partisipasi Angkatan Kerja         0
PDRB Harga Konstan                 0
Klasifikasi Kemiskinan             0
dtype: int64
```

```
Data Info:

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 514 entries, 0 to 513
Data columns (total 13 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Provinsi                    514 non-null    object
 1   Kab/Kota                    514 non-null    object
 2   P0 Persen                   514 non-null    float64
 3   Mean Lama Sekolah 15+       514 non-null    float64
 4   Pengeluaran perKapita       514 non-null    int64
 5   IPM                         514 non-null    float64
 6   Umur Harapan Hidup          514 non-null    float64
 7   Akses Sanitasi Layak        514 non-null    float64
 8   Akses Air Minum Layak       514 non-null    float64
 9   Pengangguran Terbuka        514 non-null    float64
 10  Partisipasi Angkatan Kerja  514 non-null    float64
 11  PDRB Harga Konstan          514 non-null    int64
 12  Klasifikasi Kemiskinan      514 non-null    int64
dtypes: float64(8), int64(3), object(2)
memory usage: 52.3+ KB
```

	P0 Persen	Mean Lama Sekolah 15+	Pengeluaran perKapita
count	514.000000	514.000000	514.000000
mean	12.273152	8.436615	10324.787938
std	7.458703	1.630842	2717.144186
min	2.380000	1.420000	3976.000000
25%	7.150000	7.510000	8574.000000
50%	10.455000	8.305000	10196.500000
75%	14.887500	9.337500	11719.000000
max	41.660000	12.830000	23888.000000

Data Description:

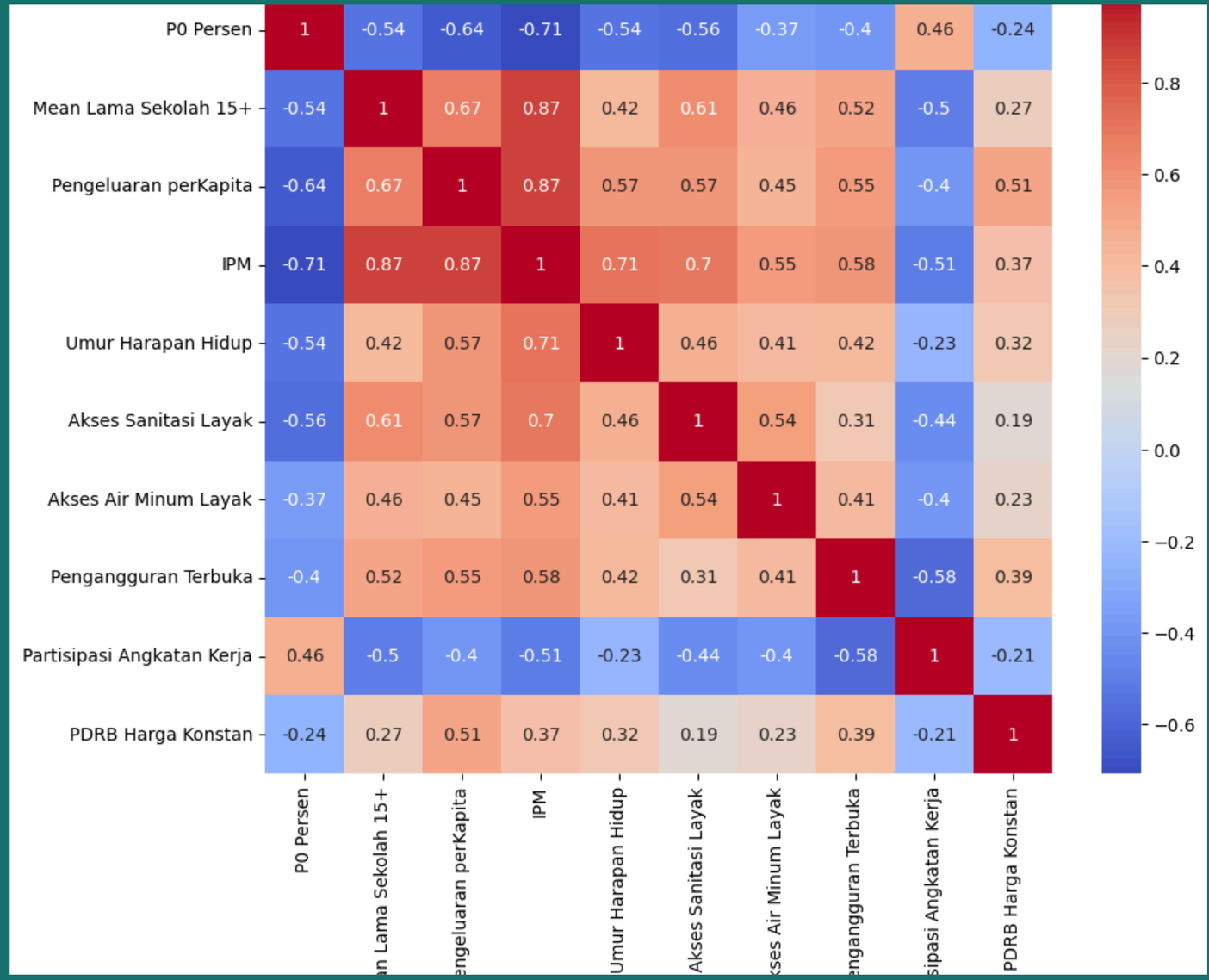
	P0 Persen	Mean Lama Sekolah 15+	Pengeluaran perKapita	IPM	Umur Harapan Hidup	Akses Sanitasi Layak	Akses Air Minum Layak	Pengangguran Terbuka	Partisi Angka K
count	514.000000	514.000000	514.000000	514.000000	514.000000	514.000000	514.000000	514.000000	514.000
mean	12.273152	8.436615	10324.787938	69.926770	69.656809	77.202237	85.136615	5.059494	69.464
std	7.458703	1.630842	2717.144186	6.497033	3.447464	18.583555	15.701658	2.636970	6.396
min	2.380000	1.420000	3976.000000	32.840000	55.430000	0.000000	0.000000	0.000000	56.390
25%	7.150000	7.510000	8574.000000	66.642500	67.387500	70.217500	79.042500	3.180000	65.070
50%	10.455000	8.305000	10196.500000	69.610000	69.975000	81.800000	89.795000	4.565000	68.955
75%	14.887500	9.337500	11719.000000	73.112500	72.042500	89.882500	96.400000	6.530000	72.342
max	41.660000	12.830000	23888.000000	87.180000	77.730000	99.970000	100.000000	13.370000	97.930

METHODOLOGY

EXPLORATORY DATA ANALYS

Insights from This Heatmap

- P0 Persen vs. IPM: Strong negative correlation (-0.71), indicating that as the poverty percentage decreases, the Human Development Index (IPM) tends to increase.
- IPM vs. Pengeluaran per Kapita: Strong positive correlation (0.87), showing that regions with higher spending per capita also tend to have higher IPM.
- Mean Lama Sekolah vs. Pengeluaran per Kapita: Moderate positive correlation (0.67), suggesting that longer education years are associated with higher per capita spending.
- Features with strong correlations, like IPM and Pengeluaran per Kapita, can be prioritized as predictors in poverty classification models.
- Weakly correlated variables may contribute less but could still provide unique information.



METHODOLOGY

PREPROCESSING DATA

- Handle missing values by filling numerical columns with the mean.
- Encode categorical columns (e.g., province/city names) using LabelEncoder.
- Scale numerical features to normalize the data using StandardScaler.

```python
# Identifikasi kolom numerik
numeric_cols = dataset.select_dtypes(include=['number'])
non_numeric_cols = dataset.select_dtypes(exclude=['number'])

# Handling missing values
dataset[numeric_cols.columns] = numeric_cols.fillna(numeric_cols.mean())
dataset[non_numeric_cols.columns] = non_numeric_cols.fillna('unknown')
```

```python
# Encoding categorical variables
label_encoders = {}
for column in non_numeric_cols.columns:
    label_encoders[column] = LabelEncoder()
    dataset[column] = label_encoders[column].fit_transform(dataset[column])

# Feature scaling
scaler = StandardScaler()
X = dataset.drop("Klasifikasi Kemiskinan", axis=1)
y = dataset["Klasifikasi Kemiskinan"]
X_scaled = scaler.fit_transform(X)
```

METHODOLOGY

SPLIT DATA

- Splitting the dataset is crucial in machine learning to evaluate how well a model can generalize to unseen data.
- The dataset is divided into two main parts:
- Training Set: Used to train the model and help it learn patterns from the data.
- Testing Set: Used to evaluate the model's performance on data it has never seen before.

Split the dataset into training (80%) and testing (20%) sets using train_test_split.

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

MODELING

Random Forest is an ensemble learning algorithm that uses multiple decision trees to make predictions.

- Handles large datasets with high dimensionality.
- Reduces the risk of overfitting by averaging results from multiple trees.
- Provides feature importance, highlighting which variables contribute most to predictions.

```
model = RandomForestClassifier(random_state=42)

# Train the model
model.fit(X_train, y_train)
```

```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

METHODOLOGY

PREDICT & EVALUATE

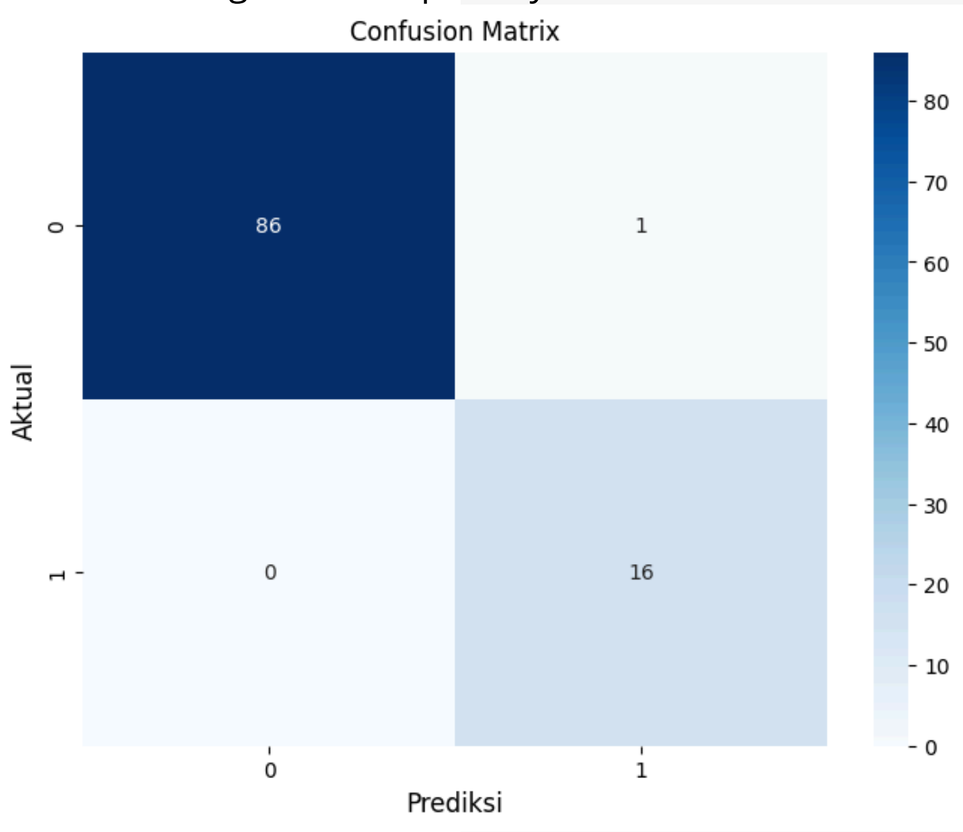
1. Classification Report:

- Accuracy: 99.03%.
- Precision, Recall, F1-Score: High performance across both classes (poverty and non-poverty).

```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.99      0.99        87
           1       0.94      1.00      0.97        16

    accuracy                           0.99       103
   macro avg       0.97      0.99      0.98       103
weighted avg       0.99      0.99      0.99       103


Random Forest Accuracy: 99.03%
```

2. Confusion Matrix:

- 86 True Negatives: Correctly predicted non-poverty.
- 16 True Positives: Correctly predicted poverty.
- 1 False Positive: Non-poverty misclassified as poverty.
- 0 False Negatives: No poverty cases misclassified.



THANK YOU

FOR YOUR NICE ATTENTION

LinkedIn
https://www.linkedin.com/in/mhabibierobbi12/

Github
https://github.com/MahirBye

January 2025