# Laporan Proyek Final Data Analyst Camp: Perbandingan Model Klasifikasi untuk Prediksi Risk Level

## 1. Pendahuluan

Dalam proyek ini, dilakukan analisis terhadap tiga metode klasifikasi untuk memprediksi Risk Level berdasarkan beberapa fitur seperti Stress Level (GSR), Sleep Hours, Anxiety Level, dan Mood Score.
Metode yang digunakan meliputi:

1. Logistic Regression (Klasifikasi Dasar)
2. Random Forest (Ensemble Learning)
3. XGBoost (Gradient Boosting)

Dokumen ini menjelaskan metode yang digunakan, implementasi model, evaluasi kinerja, serta analisis perbandingan dari ketiga pendekatan tersebut.

## 2. Metode dan Implementasi

### 2.1 Klasifikasi Menggunakan Logistic Regression

Logistic Regression adalah model klasifikasi yang menggunakan fungsi logistik (sigmoid) untuk memprediksi probabilitas dari setiap kelas.
Dalam kasus ini, digunakan *Multinomial Logistic Regression* karena variabel target memiliki lebih dari dua kelas.

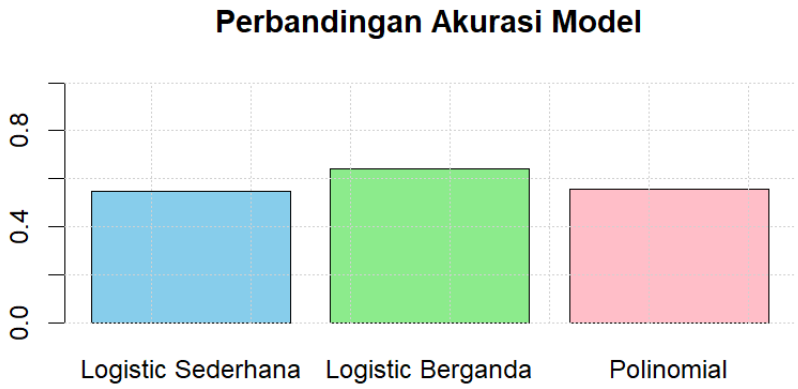
**Implementasi**
Tiga variasi Logistic Regression diuji:

- Model sederhana: Menggunakan hanya satu fitur utama (`Stress.Level..GSR.`).
- Model berganda: Menggunakan beberapa fitur tambahan (`Sleep Hours`, `Anxiety Level`, `Mood Score`).
- Model polinomial: Menambahkan fitur kuadratik untuk meningkatkan performa.

model_berganda <- multinom(Risk.Level ~ Stress.Level..GSR. + Sleep.Hours + Anxiety.Level + Mood.Score, data = train_data)

Hasil

| | Predicted | | |
|---|---|---|---|
| Actual | Low | Medium | High |
| Low | 320 | 53 | 281 |
| Medium | 33 | 314 | 322 |
| High | 166 | 220 | 1290 |




**Kelebihan**

- Cepat dan mudah diterapkan
- Dapat memberikan interpretasi yang jelas terhadap kontribusi masing-masing fitur

**Kekurangan**

- Kurang mampu menangkap hubungan non-linear antar variabel
- Kesulitan dalam memisahkan kelas yang memiliki pola serupa

**Hasil Evaluasi**

- Akurasi sekitar 64.15%.
- Model sering mengalami kesalahan dalam membedakan kelas Medium dan High.

## 2.2 Klasifikasi Menggunakan Random Forest

Random Forest adalah algoritma ensemble learning yang mengkombinasikan banyak pohon keputusan (decision trees) dan mengambil rata-rata hasil prediksi dari pohon-pohon tersebut.

**Implementasi**

```
rf_model <- randomForest(
  Risk.Level ~ .,
  data = train_data[, c(features, target)],
  ntree = 100,
  mtry = 2,
```

```
  importance = TRUE
)
```

- `ntree = 100` → Jumlah pohon dalam model.
- `mtry = 2` → Jumlah fitur yang dipilih di setiap split.
- `importance = TRUE` → Mengaktifkan analisis feature importance.

**Hasil**

```
Confusion Matrix and Statistics

          Reference
Prediction  Low Medium High
    Low     568      0  277
    Medium    0    546  311
    High     86    123 1088
```

**Kelebihan**

- Lebih akurat dibanding Logistic Regression.
- Mampu menangkap hubungan non-linear antar fitur.
- Memberikan informasi fitur paling penting dalam prediksi.

**Kekurangan**

- Lebih lambat dibanding Logistic Regression.
- Berpotensi mengalami overfitting jika tidak dilakukan tuning dengan baik.

**Hasil Evaluasi**

- Akurasi meningkat menjadi 73.42%.
- Model lebih baik dalam memprediksi kelas Low dan Medium, namun masih mengalami kesalahan pada kelas High.

## 2.3 Klasifikasi Menggunakan XGBoost

XGBoost (Extreme Gradient Boosting) adalah algoritma berbasis gradient boosting yang membangun pohon keputusan secara bertahap, di mana setiap pohon baru bertujuan untuk memperbaiki kesalahan dari pohon sebelumnya.

**Implementasi**

```
xgb_model <- xgboost(
  data = train_matrix,
```

```
  objective = "multi:softmax",
  num_class = 3,
  nrounds = 100,
  eta = 0.1,
  max_depth = 6,
  subsample = 0.8,
  colsample_bytree = 0.8,
  eval_metric = "mlogloss",
  verbose = 1
)
```

- `objective = "multi:softmax"` → Model klasifikasi multi-kelas
- `nrounds = 100` → Jumlah iterasi boosting
- `eta = 0.1` → Learning rate
- `max_depth = 6` → Kedalaman pohon keputusan

**Hasil**

```
Confusion Matrix and Statistics

          Reference
Prediction  Low Medium High
    Low     645      0  313
    Medium    0    619  344
    High      9     50 1019
```

**Kelebihan**

- Akurasi lebih tinggi dibanding Random Forest.
- Lebih cepat dan efisien untuk dataset besar.
- Mampu menangani hubungan non-linear lebih baik.

**Kekurangan**

- Membutuhkan tuning hyperparameter agar optimal.
- Lebih kompleks dibandingkan dengan Logistic Regression dan Random Forest.

**Hasil Evaluasi**

- Akurasi meningkat menjadi 76.13%, lebih tinggi dibanding Random Forest.
- Model lebih baik dalam membedakan kelas Medium dan High.

# 3. Perbandingan dan Evaluasi Model

Berdasarkan perbandingan di atas, model XGBoost memiliki akurasi tertinggi dan menunjukkan performa yang lebih baik dibandingkan Logistic Regression maupun Random Forest.

# 4. Kesimpulan dan Rekomendasi

Dari hasil penelitian ini, dapat disimpulkan bahwa:

1. Logistic Regression kurang optimal dalam memprediksi kategori yang memiliki pola kompleks.
2. Random Forest lebih baik dalam menangkap pola non-linear, tetapi masih memiliki kesulitan dalam membedakan kelas tertentu.
3. XGBoost memberikan hasil terbaik, dengan akurasi tertinggi dan kemampuan menangani hubungan antar fitur yang lebih baik.

Dokumen ini diharapkan dapat menjadi referensi dalam pemilihan model klasifikasi yang optimal untuk prediksi Risk Level berdasarkan dataset yang tersedia.

**Disusun oleh: Farrel Zandra**
**Tanggal: Sabtu, 1 Maret 2025**

GitHub repo: https://github.com/arelquack/risk-level-classification