BINUS UNIVERSITY

FINAL GROUP PROJECT DISCUSSION

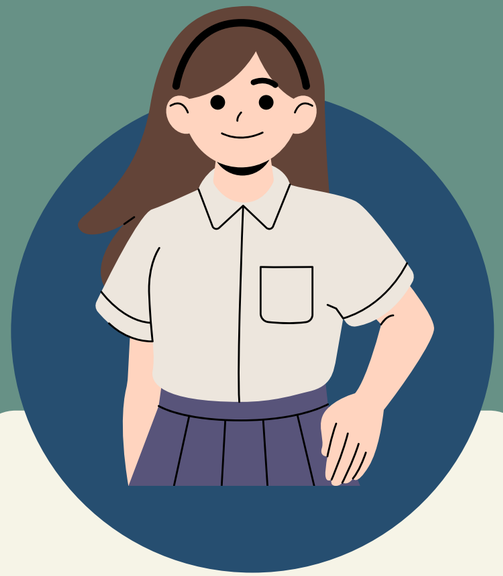
Group 2 LA08

Machine Learning

GROUP 2

Tricia Estella

2440003695



Audrey Levina

2440027921

Nadzla Andrita

2440116031

INTRODUCTION

Pada proyek ini, kami melakukan penelitian dengan menggunakan beberapa algoritma Machine Learning, dan kami beri judul yaitu Klasifikasi Indeks Pencemaran Udara di DKI Jakarta Menggunakan Machine Learning.

Kami menganalisa indeks dengan menggunakan dataset yang kami dapatkan dari kaggle.com. Dengan menggunakan empat macam algoritma, yaitu; KNN, Decision Tree, Logistic Regression, dan SVM, akhirnya kami bisa mendapatkan nilai keakurasian, serta kategori apa yang paling mempengaruhi pencemaran udara di DKI Jakarta.

OBJECTIVES

THE FIRST OBJECTIVE

Mengetahui tingkat keakurasian dalam mengklasifikasi ISPU di tiap-tiap algoritma Machine Learning

THE SECOND OBJECTIVE

Mengimplementasikan algoritma KNN, Decision Tree, Linear Regression, dan SVM secara langsung

THE THIRD OBJECTIVE

Mengetahui tahap-tahap pengolahan data dari pre-processing hingga akhir.

DEMO PROJECT

Pre-Processing
Main Processing

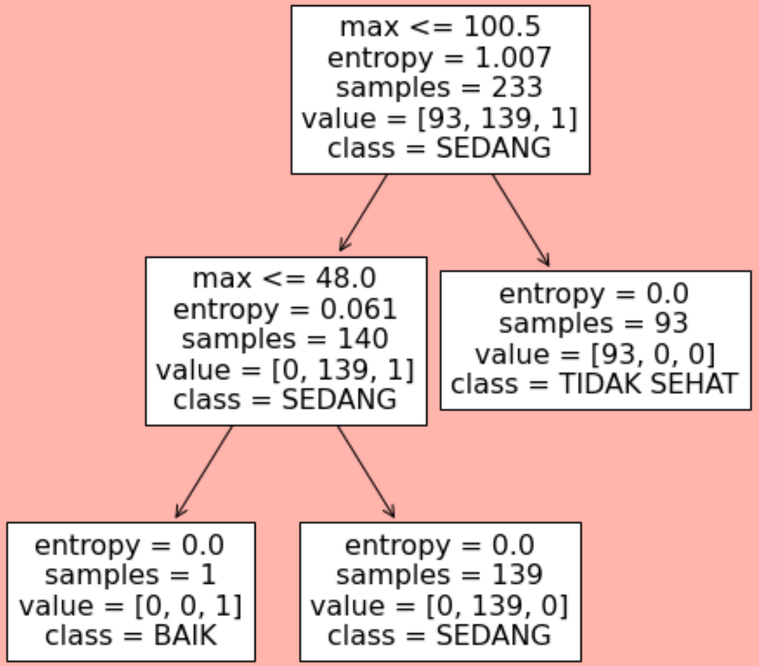
link google colab:
https://bit.ly/3n6dfjG

KNN (K-NEAREST NEIGHBOR)

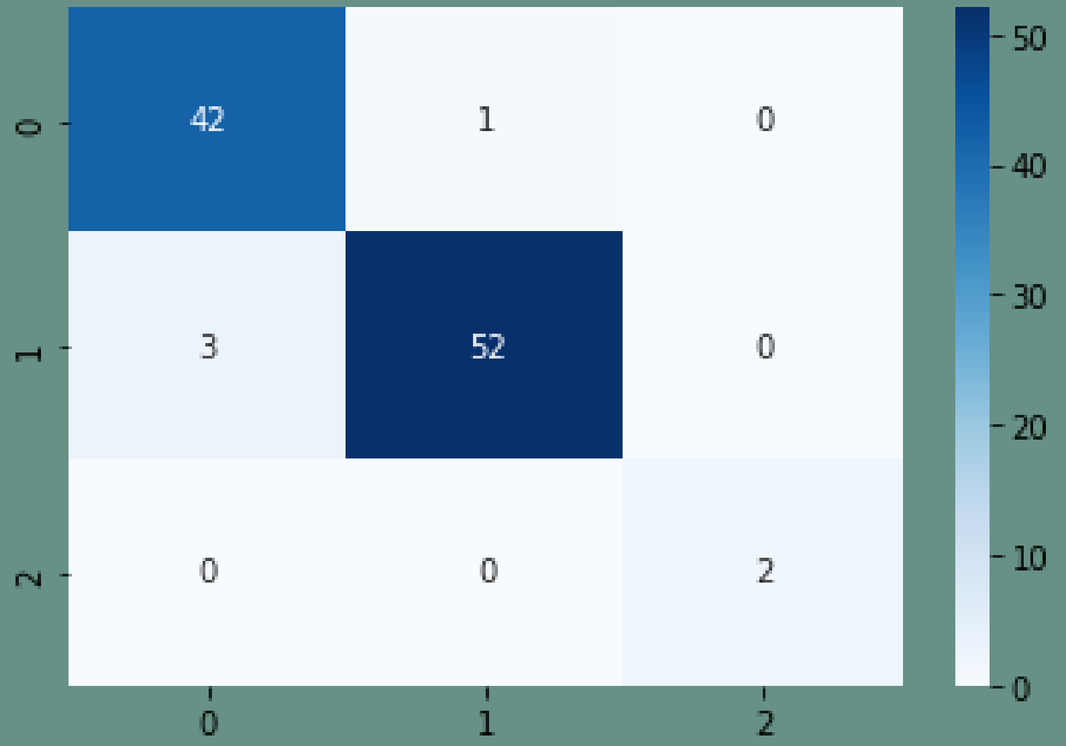
```
Train set acc:  0.9785407725321889
Test set acc:  0.94
```

Pada akhirnya, kami mendapatkan nilai akurasi 0.987 untuk train set, dan 0.96 untuk test set.
Adapun kami tidak dapat menampilkan plot untuk visualisasi KNN, dikarenakan kami menggunakan 8 buah variabel yang tidak mungkin dijadikan plot dengan 8 dimensi.

DECISION TREE



Dari graph decision tree yang terbuat, maka kami dapat melihat bahwa variabel yang digunakan hanyalah variabel 'max'.
Jika nilai max sebuah data lebih dari sama dengan 100.5, maka data tersebut akan berada di kategori tidak sehat. Jika nilai max berada di bawah sama dengan 48, maka data tersebut berada di kategori baik. Terakhir, jika sebuah nilai max berada di antara nilai 48 dan 100.5, maka data tersebut akan berada di kategori sedang.



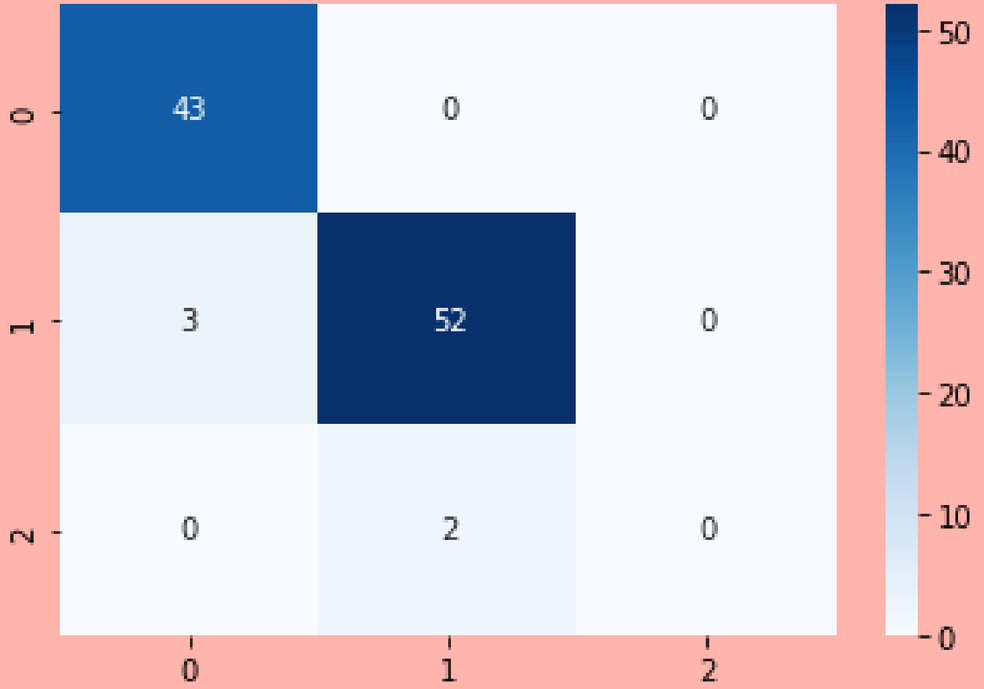
LOGISTIC REGRESSION

Hasil akurasi Train Set sangat tinggi, yaitu 100%. Hasil akurasi test set juga sangat tinggi, yaitu 96%.

Confusion matrix digunakan untuk mengevaluasi performance model klasifikasi. Hasilnya divisualisasi dalam bentuk heatmap:

- Ada 3 instansi 'tidak sehat' yang classifier kira sebagai 'sedang'
- Ada 1 instansi 'sedang' yang classifier kira sebagai 'tidak sehat'
- Selain itu, instansi lain dikelompokkan dengan benar

SUM (SUPPORT VECTOR MACHINE)



Hasil confusion matrix:
Ada 3 instansi 'tidak sehat' yang dikira sebagai 'sedang'
Ada 2 instansi 'sedang' yang dikira sebagai 'baik'
Selain itu, instansi lain dikelompokkan dengan benar.

CONCLUSIONS

Algoritma	Akurasi
K-NN	0.94
Decision Tree	0.94
Logistic Regression	0.96
SVM	0.95

Hasil akhir ini mungkin terlihat overfitting, tetapi memang data yang kami pakai untuk membuat model Machine Learning ini memang sedikit dan sederhana. Untuk kedepannya, kami memberikan saran kepada Kementerian Lingkungan Hidup dan Kehutanan Republik Indonesia, untuk menyediakan data yang lengkap dan tidak ambigu kedepannya, agar data-data tersebut bisa digunakan untuk dilakukan research dan digunakan untuk menambah ilmu masyarakat secara luas.

THANK YOU !