Belajar Mengembangkan Model Database dengan Python untuk Menjadi Analis Database / Final Projects

Objektif

Final Project 2 ini dibuat guna mengevaluasi konsep Logistic Regression dan SVM sebagai berikut:

- Mampu memahami konsep Classification dengan Logistic Regression dan SVM
- Mampu mempersiapkan data untuk digunakan dalam model Logistic Regression dan SVM
- Mampu mengimplementasikan Logistic Regression dan SVM untuk membuat prediksi

Instruksi

Final Project dikerjakan dalam format notebook dengan/atau dengan model deployment (Opsional) dengan beberapa kriteria wajib di bawah ini:

1. Machine learning framework yang digunakan adalah Scikit-Learn

2. Ada penggunaan library visualisasi, seperti matplotlib atau seaborn

3. Project dinyatakan selesai dan diterima untuk dinilai jika saat dilakukan Run All pada notebook, semua cell berhasil tereksekusi sampai akhir.

4. Isi notebook harus mengikuti outline di bawah ini:
    a. Perkenalan

        Bab pengenalan harus diisi dengan latar belakang memilih kasus, data yang digunakan (jumlah data, kelas, sumber), dan objective yang ingin dicapai.

    b. Import pustaka yang dibutuhkan

        Cell pertama pada notebook harus berisi dan hanya berisi semua library yang digunakan dalam project.

    c. Data Loading

        Bagian ini berisi proses data loading yang kemudian dilanjutkan dengan explorasi data secara sederhana.

    d. Data Cleaning

        Bagian ini berisi proses penyiapan data berupa data cleaning sebelum dilakukan explorasi data lebih lanjut. Proses cleaning dapat berupa

memberi nama baru untuk setiap kolom, mengisi missing values, menghapus kolom yang tidak dipakai, dan lain sebagainya.

e. Explorasi Data

   Bagian ini berisi explorasi data pada dataset diatas dengan menggunakan query, grouping, visualisasi sederhana, dan lain sebagainya.

f. Data Preprocessing

   Bagian ini berisi proses penyiapan data untuk proses pelatihan model, seperti pembagian data menjadi train-dev-test, transformasi data (normalisasi, encoding, dll.), dan proses-proses lain yang dibutuhkan.

g. Pendefinisian Model

   Bagian ini berisi cell untuk mendefinisikan model sampai kompilasi model. Akan lebih bagus jika didahului dengan penjelasan mengapa memilih arsitektur atau jenis model tertentu, alasan memilih nilai hyperparameter, dan hal lain yang berkaitan.

h. Pelatihan Model

   Cell pada bagian ini hanya berisi code untuk melatih model dan output yang dihasilkan.

i. Evaluasi Model

   Pada bagian ini, dilakukan evaluasi model yang harus menunjukkan bagaimana performa model berdasarkan metrics yang dipilih. Hal ini harus dibuktikan dengan visualisasi tren performa dan/atau tingkat kesalahan model. Jika memilih untuk melakukan model deployment, lanjut ke poin dibawah. Jika tidak, lanjut ke poin 5 dan 6.

j. Model Inference

   Bagian ini diisi dengan model inference, di mana model yang sudah kita latih akan dicoba pada data selain data yang sudah tersedia. Data yang dimaksud bisa berupa data buatan oleh student, ataupun data yang ada pada internet.

k. Pengambilan Kesimpulan

   Pada bab terakhir ini, harus berisi kesimpulan yang mencerminkan hasil yang didapat dengan dibandingkan dengan objective yang sudah ditulis di bagian pengenalan.

5. Notebook harus diupload dalam akun GitHub masing-masing siswa untuk selanjutnya dinilai

6. Lakukan model deployment ke Heroku.

7. Penilaian project dilakukan berdasarkan notebook dan service/API model yang sudah di-deploy (Jika melakukan model deployment).

Projects Overview

Database ini memiliki 23 atribut. Dengan data hujan harian selama 10 tahun di Australia, kolom RainTomorrow adalah target variable yang mau kita prediksi. Jika “Yes” maka besok harinya disana hujan 1mm atau lebih.

Attribute Information:

1. Date - tanggal hari itu
2. Location - lokasi, nama kota di Australia
3. MinTemp - temperatur terendah hari itu dalam celcius
4. MaxTemp - temperatur tertinggi hari itu dalam celcius
5. Rainfall - jumlah curah hujan hari itu dalam mm
6. Evaporation - jumlah evaporasi dalam mm dari Class A pan selama 24 jam sebelum jam 9 pagi hari itu
7. Sunshine - jumlah jam hari itu cerah dengan cahaya matahari
8. WindGustDir - arah kecepatan angin yang paling tinggi selama 24 jam sebelum jam 12 malam hari itu
9. WindGustSpeed - kecepatan angin yang paling tinggi dalam km/jam selama 24 jam sebelum jam 12 malam hari itu
10. WindDir9am - arah angin jam 9 pagi
11. WindDir3pm - arah angin jam 3 sore
12. WindSpeed9am - kecepatan angin jam 9 pagi dalam km/jam dihitung dari rata-rata kecepatan angin 10 menit sebelum jam 3 sore
13. WindSpeed3pm - kecepatan angin jam 3 sore dalam km/jam dihitung dari rata-rata kecepatan angin 10 menit sebelum jam 3 sore
14. Humidity9am - humiditas jam 9 pagi dalam persen
15. Humidity3pm - humiditas jam 3 sore dalam persen
16. Pressure9am - tekanan udara jam 9 pagi dalam hpa
17. Pressure3pm - tekanan udara jam 3 sore dalam hpa
18. Cloud9am - persentase langit yang tertutup awan jam 9 pagi. dihitung dalam oktas, unit ⅛, menghitung berapa unit ⅛ dari langit yang tertutup awan. Jika 0, langit cerah, jika 8, langit sepenuhnya tertutup awan.
19. Cloud3pm - persentase langit yang tertutup awan jam 3 sore
20. Temp9am - temperatur jam 9 pagi dalam celcius
21. Temp3pm - temperatur jam 3 sore dalam celcius
22. RainToday - apakah hari ini hujan: jika curah hujan 24 jam sebelum jam 9 pagi melebihi 1mm, maka nilai ini adalah 1, jika tidak nilai nya 0
23. RainTomorrow - variable yang mau di prediksi

Starting from Scratch

Kalian dapat membuat file jupyter notebook sendiri dan upload ke github ketika selesai:

1. Download Dataset yang diperlukan di sini (data Rain in Australia), lalu save ke folder /dataset.
2. Buat sebuah Notebook baru, lalu rename file menjadi “PYTN_KampusMerdeka_fp2_<nama>".
3. Bersihkan dan preproses Dataset kamu.
4. Bangun model menggunakan Logistic Regression, KNN, SVM, Naive Bayes, Decision Tree, dan Random Forest, atau teknik lainnya.
5. Pilih 1 algoritma yang kamu anggap paling sesuai lalu jelaskan mengapa.
6. Kumpulkan informasi melalui analisis kamu.

Projects Submission

Instructions

Jika kamu memilih untuk mengembangkan proyek kamu dengan komputer pribadi, kamu perlu meng-upload file kamu ke github.

Penting!

Push assignment yang telah kalian buat ke dalam akun github kalian masing-masing.

Buat sebuah file .txt dengan notepad atau code editor pilihan kalian dan masukan link repository assignment kalian dalam file .txt tersebut. Unggah file .txt tersebut pada Google Classroom.

Submission Assignment valid jika link repository assignment kalian dapat diakses untuk kemudian dinilai oleh Hacktiv8 PTP Program Code Reviewer.

Checklist

Before submitting your project, please review and confirm the following items.

- semua *bugs* sudah dibasmi
- cek rubrik dan pastikan proyek kamu sudah memenuhi semua *requirements*

Tidak usah terburu-buru dan latihan terus untuk mendapatkan semua pengetahuan baru. Kita semua bisa! 🙌

Once you have checked all these items, you are ready to submit!

Project Rubric

Proyek kamu akan dinilai oleh Hacktiv8 PTP Program Code Reviewer berdasarkan rubrik. Pastikan untuk me-review proyek kamu terlebih dahulu sebelum submit. Semua kriteria harus terpenuhi untuk mendapatkan nilai.

Code Review

Criteria	Meet Expectations
Logistic Regression	Mengimplementasikan Logistic Regression Dengan Scikit-Learn
K-Nearest Neighbors	Mengimplementasikan K-Nearest Neighbors Dengan Scikit-Learn
Support Vector Machine	Mengimplementasikan Support Vector Machine Dengan Scikit-Learn
Decision Tree	Mengimplementasikan Decision Tree Dengan Scikit-Learn
Random Forest	Mengimplementasikan Random Forest Dengan Scikit-Learn
Naive Bayes	Mengimplementasikan Naive Bayes Dengan Scikit-Learn
Confusion Matrix	Mengimplementasikan Confusion Matrix Regression Dengan Scikit-Learn
Visualization	Menganalisa Data Menggunakan Setidaknya 2 Tipe Grafik/Plot.
Preprocessing	Melakukan Preproses Dataset Sebelum Melakukan Penelitian Lebih Dalam.
Apakah Kode Berjalan Tanpa Ada Error?	Kode Berjalan Tanpa Ada Error. Seluruh Kode Berfungsi Dan Dibuat Dengan Benar.

Readability

Criteria	Meet Expectations
Tertata Dengan Baik	Semua Cell Di Notebook Terdokumentasi Dengan Baik Dengan Markdown Pada Tiap Cell Untuk Penjelasan Kode.

Analysis

Criteria	Meet Expectations
Algorithm Analysis	Student Menjelaskan Alasan Mengapa Memilih Menggunakan Algoritma Tersebut Untuk Membuat Model.

FORM FINAL PROJECT

ASPEK	RUBRIK		BOBOT	SKOR
	SKALA	DESKRIPSI		
EDA exploratory data analysis	5	tanpa error	5	...
	4	ada deskripsi hasil no2 dan no3		
	3	nomor 3, tambah mencari variability		
	2	nomor 2, mencari central tendency		
	1	melakukan data query dan grouping		
Preprocessing	5	nomor 4, Semua baris kode terdokumentasi Dengan Baik Dengan Markdown Untuk Penjelasan Kode	15	...
	4	nomor 3, menambah kolom baru dari operasi kolom		
	3	nomor 2, fix column data type		
	2	sudah tidak ada missing value atau invalid value		
	1	fix nama kolom, handle missing value namun masih ada invalid value		
Modelling	5	Mencoba model yang telah dibuat dengan data baru yang disediakan	40	...
	4	implementasi SVM dengan scikit-learn		
	3	implementasi logistic regression dengan scikit-learn		
	2	Visualisasi yang baik dengan title, label axes, legend, custom size, custom color, dan anotasi		
	1	ada visualisasi		

Analysis	5	Menarik lebih dari 1 kesimpulan	25	...
	4	Menarik Informasi/Kesimpulan Dari Keseluruhan Kegiatan yang Dilakukan		
	3	Membuat lebih dari 1 informasi		
	2	Menganalisa informasi dari model yang telah dibuat		
	1	Mendeskripsikan model yang telah dibuat		
Deployment	5	nomor 4, graded dapat melakukan inference	15	...
	4	nomor 3, tanpa error		
	3	nomor 2, melakukan deployment, ada error		
	2	nomor 1, menambahkan requirement heroku, tanpa deployment		
	1	dapat membuat flask tanpa deployment		
TOTAL SKOR				**100**