

Panduan Pelaksanaan Pertemuan

Project DTS: Machine Learning

Pendahuluan

Project merupakan sebuah praktik mandiri yang diberikan ke peserta. Pelaksanaan project ditujukan untuk menilai seberapa jauh pemahaman yang dimiliki oleh peserta mengenai teori yang telah didapatkannya. Sebuah studi kasus akan diberikan kepada peserta dan kemampuan analisis serta pemecahan masalah peserta akan diuji secara langsung.

Terdapat dua buah pelaksanaan project yang harus dilakukan peserta, keduanya berlangsung di hari sebelum ujian (pertengahan dan akhir). Project pertama akan disebut sebagai Project Machine Learning, sedangkan project kedua akan disebut dengan Project Deep Learning.

Peraturan

Berikut merupakan beberapa peraturan dalam pelaksanaan project:

1. Pelaksanaan Project:
 - a. Project Machine Learning dilakukan secara individu dan dilaksanakan 1 hari.
 - b. Project Deep Learning dilakukan secara kelompok dan dilaksanakan selama 2 hari. Selain berupa studi kasus, presentasi dilakukan kelompok di hari ke-2.
2. File yang dikumpulkan:
 - a. Project Machine Learning dikerjakan menggunakan Jupyter Notebook (Google Colab) dan disimpan kedalam file dengan format .ipynb. Project dikumpulkan ke Instruktur/Asisten.
 - b. Khusus untuk Project Deep Learning, selain file berformat .ipymb, file PowerPoint dengan format .ppt/.pptx juga dikumpulkan ke Instruktur/Asisten.
3. Project dilaksanakan secara mandiri tanpa arahan dari Instruktur/Asisten. Seluruh informasi permasalahan yang harus diselesaikan terjabarkan dalam studi kasus.
4. Peserta diperbolehkan mencari informasi dokumentasi library yang akan ia pakai menggunakan koneksi Internet.
5. Khusus untuk Project Machine Learning, peserta tidak boleh mencari penyelesaian masalah yang tersedia secara online. Instruktur/Asisten harus mengawasi pekerjaan yang dilakukan oleh peserta.

Persiapan

Terdapat beberapa hal yang perlu disiapkan untuk keberlangsungan pelaksanaan project, antara lain:

1. Koneksi internet dan akses menuju Google Colabs

Koneksi internet diwajibkan tersedia agar peserta dapat mengerjakan projectnya dengan Google Colabs. Koneksi internet juga memungkinkan peserta untuk mencari informasi dokumentasi library yang ia pakai.

2. Project Description

Project Description merupakan dokumen yang berisi deskripsi dari project yang harus diselesaikan oleh peserta. Terdapat beberapa studi kasus, permasalahan, dan informasi data dan algoritma yang akan digunakan ada didalamnya. Project Description terlampir dan boleh di bagikan untuk para peserta, boleh softcopy maupun hardcopy.

3. Dataset

Setiap studi kasus akan menggunakan dataset yang berbeda-beda. Ada beberapa dataset yang berukuran kecil ada juga dataset yang berukuran sangat besar (khususnya dataset untuk menunjang Deep Learning). Dataset berukuran besar harus sudah dipersiapkan oleh panitia dan didistribusikan secara offline paling lambat 1 hari sebelum hari pelaksanaan project.

Lampiran Project Description: ML

Klasifikasi Bunga Iris

Iris adalah tanaman berbunga dengan bunga yang cukup mencolok karena warnanya. Bunga Iris ini diobservasi dan dikumpulkan data-datanya oleh biolog Inggris pada tahun 1936. Tujuan pengumpulan data bunga Iris ini untuk menghitung variasi morfologis bunga iris dari tiga spesies yang terkait, yakni Iris Sentosa, Iris Virginica, dan Iris Versicolor. Empat fitur utama diukur dari masing-masing sampel: panjang dan lebar dari sepal dan kelopak, dalam sentimeter (lihat gambar dibawah).



Gambar 1 Tiga Spesies Bunga Iris

Berdasarkan data-data yang tersedia dalam dataset, lakukan klasifikasi terhadap ketiga bunga spesies bunga Iris tersebut. Gunakan salah satu algoritma clustering dan buat model classifier yang terbaik yang dapat memisahkan ke-tiga kelompok bunga tersebut.

Dataset yang digunakan: iris.csv

Deskripsi field dalam dataset:

Id	SPL-SPW-PTL-PTW (CM)
SepalLengthCm	Length of the sepal (in cm)
SepalWidthCm	Width of the sepal (in cm)
PetalLengthCm	Length of the petal (in cm)
PetalWidthCm	Width of the petal (in cm)
Species	Species name

Survivabilitas Korban Kapal RMS Titanic

Tenggelamnya RMS Titanic adalah salah satu kejadian kapal karam paling terkenal dalam sejarah. Pada 15 April 1912, selama pelayaran perdananya, Titanic tenggelam setelah bertabrakan dengan gunung es, menewaskan 1502 dari 2.224 penumpang dan awak.



Gambar 2 Ilustrasi Gambar Kapal Titanic

Salah satu alasan mengapa kapal karam itu menyebabkan hilangnya nyawa adalah karena tidak ada cukup sekoci untuk penumpang dan awak. Meskipun ada beberapa unsur keberuntungan dalam bertahan dari tenggelamnya, beberapa kelompok orang lebih mungkin bertahan daripada yang lain, seperti wanita, anak-anak, dan kelas atas.

Dataset berisi informasi-informasi penumpang kapal RMS Titanic. Lakukan analisa penumpang seperti apa yang memiliki kemungkinan untuk bertahan hidup. Gunakan salah satu algoritma klasifikasi untuk memprediksi penumpang mana yang selamat dari kejadian itu.

Dataset yang digunakan: titanic.csv

Deksripsi field dalam dataset:

PassengerId	1
Pclass	1
Name	the name of the passenger
Sex	gender of the passenger
Age	age of the passenger
SibSp	of siblings / spouses aboard the Titanic
Parch	of parents / children aboard the Titanic
Ticket	Ticket number
Fare	Passenger fare
Cabin	Cabin number
Embarked	Port of Embarkation

Prediksi Harga Rumah di King County, USA

King County adalah sebuah county (kabupaten) di USA yang terletak di negara bagian Washington DC. King County adalah salah satu dari tiga county Washington yang termasuk dalam area metropolitan Seattle-Tacoma-Bellevue. Pada 2011, King County adalah county berpenghasilan tertinggi ke-86 di Amerika Serikat.

King County memiliki bangunan rumah yang memiliki beberapa fitur utama, seperti: jumlah kamar tidur, jumlah kamar mandi, luas area rumah, dll. Tentu fitur-fitur utama tersebut akan mempengaruhi harga rumah tersebut. Survey dilakukan dan datanya, termasuk harga, dikumpulkan dan tersimpan dalam sebuah dataset.

Berdasarkan dataset tersebut, lakukan analisa pengaruh fitur-fitur rumah terhadap harga rumah tersebut. Menggunakan algoritma prediksi regression, buatlah sebuah model yang merepresentasikan fenomena tersebut dan gunakan model tersebut untuk memprediksikan harga beberapa rumah lainnya.

Dataset yang digunakan: kc_house_data.csv

Deskripsi field dalam dataset:

id	a notation for a house
date	Date house was sold
price	Price is prediction target
bedrooms	Number of Bedrooms/House
bathrooms	Number of bathrooms/House
sqft_living	square footage of the home
sqft_lot	square footage of the lot
floors	Total floors (levels) in house
waterfront	House which has a view to a waterfront
view	Has been viewed
condition	How good the condition is (Overall)
grade	overall grade given to the housing unit, based on King County grading system
sqft_above	square footage of house apart from basement
sqft_basement	square footage of the basement
yr_built	Built Year
yr_renovated	Year when house was renovated
zipcode	zip
lat	Latitude coordinate
long	Longitude coordinate
sqft_living15	Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area