Data Preparation

Kecerdasan buatan dalam pengembangannya tidak dapat dipisahkan dari data. Data merupakan sekumpulan fakta yang dibuat dengan kata-kata, kalimat, simbol, angka, dan lainnya. Data tidak dapat dipisahkan dari pengembangan kecerdasan buatan karena nantinya data akan digunakan sebagai "makanan" bagi model AI untuk dapat tumbuh berdasarkan data yang diberikan. Setiap objek di dunia ini memiliki data baik yang kita sadari atau tidak, seperti udara yang memiliki data suhu, kelembapan, konsentrasi, dan lain-lain. Pengumpulan data-data tersebut dalam AI Project Cycle disebut Data Acquisition. Setelah data kita dapatkan, kita masih perlu memahami kumpulan data-data tersebut, untuk memahami data dalam jumlah yang sedikit mungkin dapat hanya dibaca saja, namun bagaimana dengan data yang memiliki banyak parameter? Untuk mempermudah pemahaman data tersebut kita dapat dibantu dengan menggunakan visualisasi data, pada Project AI Cycle hal ini disebut dengan Data Exploration. Pemahaman tentang data kita perlukan agar kita dapat memilih model AI dengan tepat. Setelah kita paham terhadap data yang kita miliki, berarti kita akan tahu jika data tersebut memiliki data-data tak wajar ataupun data-data yang tidak diperlukan dalam pengembangan, oleh karena itu kita perlu membersihkan data tersebut agar dapat meningkatkan kualitas model yang dibuat, hal ini digambarkan dalam istilah Garbage In Garbage Out (GIGO) yang berarti kalau input sampah maka keluarannya juga sampah. Untuk memahami lebih lanjut mengenai pengembangan AI, mari pertama-tama kita menjadi Data Scientist yang baik dengan berkenalan dengan data.

Jenis-jenis Data

Dalam memahami data, jenis-jenis data memainkan peran yang sangat penting agar data dapat dipahami dengan baik dan kesimpulan asumsi terhadap data tersebut dapat diambil dengan benar. Jenis data yang umum digunakan adalah Data Kuantitatif, Data Kualitatif, Data Interval, Data Rasio, dan Data Ordinal. Namun, dalam pembelajaran pengembangan AI dasar kita cukup memahami Data Kuantitatif dan Data Kualitatif.

1. Data Kuantitatif

   Data Kuantitatif merupakan data yang digunakan untuk menyatakan besaran, jumlah, atau jangkauan tertentu. Data Kuantitatif sering kali digunakan untuk menyatakan besaran-besaran fisis seperti tinggi, berat, suhu, dan lain-lain. Data Kuantitatif ini dapat dibagi lagi menjadi dua jenis yakni:

   - Data Diskrit

     Pada dasarnya data diskrit merupakan data dengan nilai yang tidak dapat dibagi menjadi lebih kecil lagi, sebagai contoh jumlah orang dalam suatu ruangan yang hanya dapat dihitung masing-masing orang, ataupun jumlah salah atau benar suatu tes yang hanya dapat dinyatakan salah dan benar saja. Kedua contoh tersebut tidak dapat dinyatakan ke dalam tingkat yang lain.
   - Data Kontinu

     Berbeda dengan data diskrit, data kontinu mewakili suatu nilai yang dapat dibagi ke tingkatan lain baik lebih besar atau lebih kecil. Seperti besaran berat dapat diukur dalam kilogram atau dengan besaran yang lebih kecil atau teliti seperti gram.
2. Data Kualitatif

   Data Kualitatif merupakan data yang didefinisikan sebagai data yang mendekati serta mencirikan, dan dapat diamati. Tidak seperti Data Kuantitatif yang dapat dihitung, data kualitatif

bukanlah data yang dapat dihitung karena data ini bersifat non-numerik. Sebagai contoh untuk data kualitatif adalah warna sebuah motor yang dapat dinyatakan dengan warna kuning, hitam, merah, atau warna lainnya.

Pembersihan Data Buruk

Seperti yang sebelumnya telah disebutkan bahwa Garbage In Garbage Out (GIGO), maka kita perlu "membersihkan" atau mengolah data yang kita miliki agar data yang kita miliki menjadi lebih berarti bagi model AI yang kita buat. Pada dasarnya persiapan data seperti membersihkan data, dapat dilakukan menggunakan aplikasi seperti Microsoft Excel, Google Sheet, atau aplikasi sejenis lainnya. Namun, dalam pembelajaran kali ini kita akan mempersiapkan data menggunakan Python melalui Google Colab, untuk notebook yang digunakan dapat diakses melalui tautan berikut.

https://colab.research.google.com/drive/1dW3WaAmrNAC-jGAINcRGUFP4vdvH70nV#scrollTo=EqvwdW9wpR6e

Data Visualization

Pemahaman data pada dasarnya bisa dilakukan hanya dengan membaca data yang kita miliki saja, namun untuk data dengan dimensi yang besar, dalam kata lain banyak informasi di dalamnya, menjadi lebih sulit untuk dipahami. Oleh karena itu, untuk mempermudah kita dalam memahami data, kita dapat memvisualisasikannya sebagai grafik atau yang lainnya agar dapat lebih mudah untuk dipahami. Visualisasi data dapat dilakukan dalam berbagai aplikasi seperti Tableau, Google Sheet, atau aplikasi sejenisnya. Namun, dalam materi ini kita akan melakukan visualisasi data menggunakan Google Colab dengan notebook yang dapat diakses melalui tautan berikut.

https://colab.research.google.com/drive/1dW3WaAmrNAC-jGAINcRGUFP4vdvH70nV#scrollTo=C1HWuC23GWoJ

AI Modelling

Dalam pemodelan AI kita dapat menggunakan library dalam pengembangan masing-masing modelnya, bahkan terdapat beberapa library yang dapat mempersingkat prosesnya untuk banyak model dalam sekali eksekusi, library tersebut adalah PyCaret. Walaupun seluruh proses telah dipersingkat dengan adanya PyCaret, namun kita masih perlu melakukan evaluasi performa model terhadap data, hal ini pun menjadi kekurangan dari PyCaret karena seluruh proses telah dikemas menjadi perintah yang singkat sehingga terdapat beberapa parameter atau hyperparameter yang tidak dapat diubah sesuai kebutuhan, meskipun dapat di-tuning dengan fungsi model_tune. Walau demikian kita tetap perlu memahami beberapa istilah dalam evaluasi performa machine learning.

PyCaret

Seperti yang sebelumnya telah dijelaskan, kita akan melakukan pemodelan AI menggunakan library PyCaret. Terdapat berbagai macam hal yang perlu dijelaskan dalam penggunaan PyCaret, mencakup mempersiapkan dataset, mempersiapkan lingkungan PyCaret, melakukan pembuatan dan komparasi model, tuning model, dan export model menjadi file pickle serta memuat file pickle untuk digunakan kembali sebagai model. Hal-hal tersebut akan dijelaskan lebih rinci melalui link di bawah ini.

https://colab.research.google.com/drive/1WH5HZeekeZUzHV77AOaoCVxZ5tjpeD8x?usp=sharing

Evaluasi Model AI

Umumnya dalam melakukan evaluasi model kita akan membuat Confusion Matrix. Berdasarkan Confusion Matrix yang didapatkan, kita akan dapat menentukan Accuracy, Precission, dan Recall. Kita perlu memahami Confusion Matrix terlebih dahulu yang merupakan alat analitik prediktif yang membandingkan dan menampilkan nilai sebenarnya dengan nilai hasil prediksi model yang digunakan. Confusion Matrix dapat disusun dengan tabel yang memiliki nilai True Positive (TP), False Positive (FP), False Negative (FN), dan True Negative (TN) yang digambarkan dengan ilustrasi di bawah ini.

1. True Positive (TP) : Jumlah data yang bernilai Positif dan diprediksi benar sebagai Positif.
2. False Positive (FP) : Jumlah data yang bernilai Negatif tetapi diprediksi sebagai Positif.
3. False Negative (FN) : Jumlah data yang bernilai Positif tetapi diprediksi sebagai Negatif.
4. True Negative (TN) : Jumlah data yang bernilai Negatif dan diprediksi benar sebagai Negatif.

Setelah kita memahami Confusion Matrix kita dapat melanjutkan langkah untuk memahami beberapa hal berikut:

Accuracy

Akurasi merupakan jumlah data bernilai positif yang diprediksi positif dan data negatif diprediksi negatif dibagi dengan total data yang ada, sehingga rumusnya adalah

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Recall

Recall adalah peluang kasus dengan kategori positif yang dengan tepat diprediksi positif, yang dapat dihitung dengan

$$Recall = \frac{TP}{TP + FN}$$

Precission

Precision adalah peluang kasus yang diprediksi positif yang pada kenyataannya termasuk kasus kategori positif, sehingga dapat dituliskan sebagai

$$Precission = \frac{TP}{TP + FP}$$

F1-Score

Juga dikenal sebagai F-Measure, merupakan nilai yang didapatkan dengan menggunakan hasil Precission dan Recall yang dapat dihitung dengan

$$F1 = \frac{2 \times Recall \times Precission}{Recall + Precission}$$