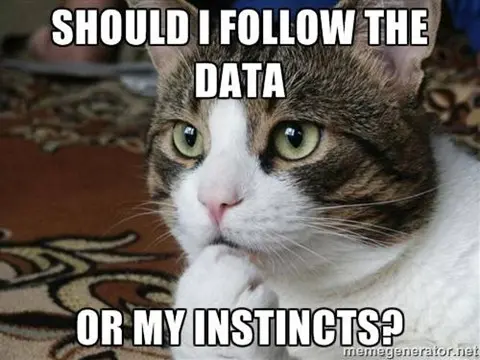
**Tugas Besar 2 IF3070 Dasar Inteligensi Artifisial**

**Implementasi Algoritma Pembelajaran Mesin**

Dipersiapkan Oleh Tim Asisten Lab AI ‘21

Versi: 1.0 07/12/2024

Versi: 1.1 7/12/2024



Deadline: Minggu, 22 Desember 2024 23.59 WIB

# Tujuan

Tugas Besar 2 pada kuliah IF3070 Dasar Inteligensi Buatan bertujuan untuk memberikan pengalaman langsung kepada peserta kuliah dalam menerapkan algoritma pembelajaran mesin pada permasalahan nyata.

# Spesifikasi

**Pembelajaran mesin** merupakan salah satu cabang dari kecerdasan buatan yang memungkinkan sistem untuk belajar dari data dan membuat prediksi atau keputusan tanpa diprogram secara eksplisit.

Pada tugas ini, Anda diminta untuk mengimplementasikan algoritma pembelajaran mesin yang telah kalian pelajari di kuliah, yaitu **KNN dan Gaussian Naive-Bayes** pada dataset **PhiUSIIL Phising URL Dataset**. Rincian spesifikasi untuk tugas besar 2 dapat dilihat sebagai berikut:

1. Implementasi KNN *from scratch*.
   1. Minimal bisa menerima 2 input parameter
      1. Jumlah tetangga
      2. Metrik jarak antar data point. Minimal dapat menerima 3 pilihan, yaitu Euclidean, Manhattan, dan Minkowski
2. Implementasi Gaussian Naive-Bayes *from scratch*.
3. Implementasi algoritma poin 1-2 menggunakan *scikit-learn*. Bandingkan hasil dari algoritma *from scratch* dan algoritma *scikit-learn*.
4. Model harus bisa di-save dan di-load. Implementasinya dibebaskan (misal menggunakan .txt, .pkl, dll).
5. [Bonus] Kaggle Submission pada link [berikut](https://www.kaggle.com/t/e984b521ddb84d14901e559d3e2ed7df).

Implementasi KNN dan Gaussian Naive-Bayes yang *from scratch* bisa dalam bentuk kelas-kelas (class KNN, dst.) yang nantinya akan di-import ke notebook pengerjaan. Untuk implementasi *from scratch*, *library* yang boleh digunakan adalah untuk perhitungan matematika saja seperti numpy dan sejenisnya.

Asisten telah menyediakan notebook [berikut](https://drive.google.com/file/d/1DWLLx6vAHnGq42BdJ_GFFCKmFwohutBo/view?usp=sharing) untuk Anda lengkapi, dan deskripsi lengkap mengenai dataset dapat dilihat sebagai berikut:

## **Deskripsi Dataset**

[PhiUSIIL Phishing URL Dataset](https://archive.ics.uci.edu/dataset/967/phiusiil+phishing+url+dataset) adalah dataset yang terdiri dari deskripsi suatu URL dan juga fitur-fitur yang terkait dengan URL tersebut beserta label URL legitimate dan URL phishing. Label 1 merupakan label untuk URL legitimate, sementara label 0 merupakan label untuk URL phishing. Sebagian besar URL yang dianalisis saat membangun dataset ini adalah URL terbaru. Fitur-fitur diekstraksi dari source code halaman web dan URL. Dataset dapat diakses pada tautan [berikut](https://www.kaggle.com/competitions/tugas-besar-2-if-3070/data).

# Bonus

Untuk bonus, nilai diberikan berdasarkan ranking *leaderboard* **Kaggle** yang dirincikan sebagai berikut:

* Rank 1-3 = 10 poin
* Rank 4-5 = 5 poin
* Rank 6-10 = 3 poin

Dalam leaderboard, gunakan nama kelompok. Identifikasi dilakukan berdasarkan nama kelompok, ~~jadi cukup 1 orang saja yang berada dalam tim Kaggle~~. Hasil prediksi di kaggle harus **reproducible**, sehingga notebook yang dikumpulkan harus bisa menghasilkan nilai akhir yang sama dengan submisi kaggle. Jika tidak sama, maka akan **didiskualifikasi** dari leaderboard. **Model yang boleh digunakan hanya KNN dan Naive Bayes yang diimplementasi from scratch.**

# Kelompok

Pembagian kelompok ditentukan sendiri oleh mahasiswa dengan mengisi [sheets kelompok](https://docs.google.com/spreadsheets/d/1lEqKoWnaYu54FnjlPgeEnIZRaT0vjSp-fUyT-v3j6w0/edit?usp=sharing) berikut ini dengan 1 kelompok terdiri dari **1-3 kelompok Tugas Kecil 2** dengan **maksimal anggota sebanyak 5 orang**. Batas waktu pengisian kelompok adalah **Senin, 9 Desember 2024 pukul 23:59 WIB**. Setelah waktu yang ditentukan, mahasiswa yang belum mengisi sheets kelompok akan diacak.

# QnA

Pertanyaan dapat ditanyakan pada [**link QnA**](https://docs.google.com/spreadsheets/d/1vdcs6_H2dzAcGRpMfuVdMeMgwbOueonr9H8Ktf1EsM0/edit?gid=560206619#gid=560206619) berikut. Pastikan pertanyaan yang ditanyakan tidak berulang.

# Aturan

Terdapat beberapa hal yang harus diperhatikan dalam pengerjaan tugas ini, yakni:

1. Jika terdapat hal yang tidak dimengerti, silahkan ajukan pertanyaan kepada asisten melalui **link QnA** yang telah diberikan di atas. Pertanyaan yang diajukan secara personal ke asisten **tidak akan dijawab** untuk menghindari perbedaan informasi yang didapatkan oleh peserta kuliah.
2. Dilarang melakukan **plagiarisme,** **menggunakan AI dalam bentuk apapun secara tidak bertanggungjawab, dan melakukan kerjasama antar kelompok**. Pelanggaran pada poin ini akan menyebabkan pemberian **nilai E** pada setiap anggota kelompok.

# Deliverables

* Tugas dikumpulkan dalam bentuk link ke *repository* GitHub yang **minimal** berisi beberapa hal berikut (boleh ditambahkan jika dirasa perlu):
  + Folder **src**, digunakan untuk menyimpan source code
  + Folder **doc,** digunakan untuk menyimpan laporan dalam bentuk **.pdf** yang terdiri atas komponen berikut:
    - Cover
    - Penjelasan singkat implementasi KNN.
    - Penjelasan singkat implementasi Naive-Bayes.
    - Penjelasan tahap cleaning dan preprocessing yang dilakukan beserta dengan alasannya.
    - Perbandingan hasil prediksi dari algoritma yang diimplementasikan dengan hasil yang didapatkan dengan menggunakan pustaka. Jelaskan insight yang kalian dapatkan dari perbandingan tersebut.
      * Perbandingan hasil dapat menggunakan metrics yang sesuai dengan permasalahan yang ada.
    - Kontribusi setiap anggota dalam kelompok.
    - Referensi
  + **README.md**, yang berisi deskripsi singkat repository, cara setup dan run program, dan pembagian tugas tiap anggota kelompok.
* Pengumpulan dilakukan melalui form dengan tautan sebagai [berikut](https://docs.google.com/forms/d/e/1FAIpQLSfbeemLt-OkU_JpDRqbeAY_BWs6JZRdZzj7gbUtuePUGo71Cw/viewform?usp=sharing).
* Batas akhir pengumpulan adalah hari **Minggu, 22 Desember 2024 23.59 WIB**. Tugas yang terlambat dikumpulkan tidak akan diterima.
* Pengumpulan dilakukan oleh NIM terkecil.

# Referensi

* <https://archive.ics.uci.edu/dataset/967/phiusiil+phishing+url+dataset>
* <https://www.sciencedirect.com/science/article/abs/pii/S0167404823004558?via%3Dihub>
* <https://scikit-learn.org/1.5/modules/neighbors.html>
* <https://scikit-learn.org/1.5/modules/naive_bayes.html>