**Nama:** Limas Jaya Akeh **| Kelas:** LB-01 | **NIM:** 2201763240 | Machine Learning

**Session 1**

**Introduction to Machine Learning**

**Review Question**

1. How would you define Machine Learning?

2. Would you frame the problem of spam detection as a supervised learning problem or an unsupervised learning problem?

3. What are the two most common supervised tasks?

4. Can you name four common unsupervised tasks?

**Answers**

**1. How would you define Machine Learning?**

Machine Learning adalah bagian dari Artificial Intelligence yang terfokus pada mesin atau system yang dapat belajar dari data dan menjadi cerdas, terutama dalam melakukan hal-hal yang memerlukan tingkat keahlian yang rendah (low-skill tasks). Machine Learning memiliki tahap-tahap umum yaitu, *Data Gathering, Reprocessing, Feature Extraction, Create ML Algorithm*

*Data Gathering* adalah tahap pengumpulan data dari topik yang sudah ditentukan. Data yang dikumpulkan harus relevan terhadap masalah yang ingin dipecahkan, tahap ini dapat dilakukan dengan banyak cara, seperti survey, questioner, interview, sensus, Expert advisory, dan lain-lainnya.

*Reprocessing* adalah tahap persiapan, filtering atau pembersihan data agar relevan sesuai kebutuhan. Data ini di-load dalam mesin untuk persiapan Machine Learning. Data akan dibersihkan agar tidak ada satupun yang kotor (tidak ada NULL / data tidak ada), di normalisasi, dan dibelah menjadi 2 set, Training set dan Validation set.

*Feature Extraction* adalah tahap pengambilan fitur, atau sesuatu yang dapat membedakan tiap atribut dalam model, sebagai contohnya apa yang perlu diperhatikan dalam Image Recognition apabila ingin meng-identifikasi antara kucing, anjing, dan rubah? Mungkin dari bentuk ekor, dari warna yang mencolok seperti warna bulu, dan bentuk hidung.

*Create ML Algorithm* adalah tahap terakhir dalam Machine Learning dimana akan dipilih algoritma yang paling cocok untuk menerapkan dan mencari solusi. Algoritma seperti K-Nearest Neighbor, dan Random Forest.

**2. Would you frame the problem of spam detection as a supervised learning problem or an unsupervised learning problem?**

Spam Detection menurut saya lebih mengarah kepada **Supervised Learning problem**, tetapi dapat juga menggunakan Unsupervised Learning problem.

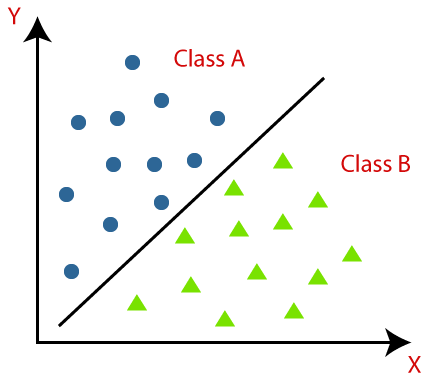
Saat menggunakan Supervised Learning, mesin akan diberikan training set berupa contoh-contoh not-spam, dan contoh-contoh yang merupakan spam; dengan ini, mesin dapat mengetahui karakteristik yang berupa spam seperti kata-kata kunci yang sering digunakan oleh spammer, dan kemudian mencoba untuk menandai seberapa banyak kata kunci tersebut muncul dalam satu mail. Hal ini memungkinkan mesin untuk mengenal dan dapat filter secara cepat mail yang berupa spam dan mana yang bukan merupakan spam.

Ketika menggunakan Unsupervised Learning, mesin diberikan training set yang tidak dilabel, kemudian dapat menggunakan clustering untuk membagi set yang ada menjadi suatu grup. Setidaknya dengan pembagian yang ada, dapat diketahui bahwa salah satu grup merupakan ciri-ciri mail berupa spam.

**3. What are the two most common supervised tasks?**

2 Tugas yang paling sering digunakan sebagai aplikasi dalam Supervised Task adalah Classification dan Regression.

- Classification



Classification adalah tugas untuk mengidentifikasi dan mengetahui suatu data baru masuk ke dalam kategori dan class yang ada berdasarkan atribut yang ditetapkan, contoh klasifikasi seperti Spam Filtering, dimana tiap e-mail dikategorikan menjadi “Spam” atau “Not Spam”. Beberapa Model Classification adalah Random Forest, Multilayer Perceptron, Naïve Beyes.

- Regression  
 Regression adalah tugas untuk memprediksi suatu nilai atau value berdasarkan variable independen dan mencari hubungan antara variable. Beda dengan Classification, jika Classification outputnya berupa kategori atau diskrit, Regression mengeluarkan output dalam bentuk numerical atau continuous. Contoh Regression seperti mencari dan memprediksi pertumbuhan penduduk tiap tahunnya.

**4. Can you name four common unsupervised tasks?**

Unsupervised Learning seringkali digunakan dalam hal-hal yang berhubungan dengan Clustering atau pembagian kelompok, seperti:

- Clustering

A picture containing light, star

Description automatically generated

Clustering adalah sebuah task untuk mengelompokkan objek yang memiliki karakteristik yang mirip menjadi sebuah kelompok atau grup, Clustering masuk kedalam Unsupervised Learning karena mesin tidak perlu tahu label dari tiap grup, hanya perlu mengetahui fitur-fitur yang membedakan antar kelompok.

- Dimensionality Reduction

PCA adalah salah satu teknik Dimensionality Reduction, yaitu sebuah Teknik untuk menyederhanakan suatu data, dan kemudian mereduksi dimensi suatu data tanpa mengurangi karakteristik data secara signifikan, PCA berusaha agar konten data sebisa mungkin tidak berubah saat di reduksi dan memiliki makna yang sama. PCA termasuk unsupervised task karena hasil yang dikeluarkan berupa data itu sendiri dan tidak perlu memerlukan label.

- Finding Association Rules

Association Rules adalah sebuah analisis untuk mengetahui hubungan antar objek, variable, dan sebagainya. Association Rules sering digunakan untuk memprediksi pembelian konsumen, sebagai contoh, konsumen suka membeli barang X dan Y secara bersamaan. Karena kita tidak mengetahui bagaimana relasi tiap barang sebelumnya, maka termasuk kedalam Unsupervised Learning

- Anomaly Detection

Deteksi anomaly adalah identifikasi dan observasi sesuatu secara unik maupun sesuatu yang mencurigakan karena cukup berbeda dari mayoritas data yang ada, deteksi anomaly dapat dilakukan oleh mesin ketika sesuatu terjadi secara tidak wajar dalam suatu periode.

A close up of a device

Description automatically generated

Salah satu contohnya adalah deteksi tindak kejahatan fraud dalam kartu kredit, apabila suatu transaksi yang tidak di-authorized dari pengguna dan terjadi dalam suatu waktu secara cepat, dan pengguna jarang menggunakan kartu kredit, mesin dapat menandai dan mencoba untuk mengkonfirmasi apakah kartu kredit digunakan oleh pengguna atau digunakan oleh pihak lainnya.

**Case Study**

1. Analyse the data and give a suggestion of type of machine learning system that can be applied.
2. Try in Google Collaboratory to explore the data, for instance:

- Find minimum or maximum of a value

- Eliminate null value

- Replace null value to certain value

**Answer**

1. Tipe Machine Learning yang cocok untuk dataset Airbnb tersebut adalah Supervised Learning, Online Learning, dan Model-based Learning

Menggunakan Supervised Learning karena data yang tersedia perlu dilabel, seperti label melalui kategori tipe room\_type. Supervised Learning memudahkan mesin belajar karena dapat mengetahui tipe ruangan apa saja beserta spesifikasinya berdasarkan alamat, harga, dan review-review sebelumnya.

Menggunakan Online Learning karena komputasi yang diperlukan sangat besar, terutama apabila dataset yang ada sangat besar, dan data flow secara kontinu / masuk terus-menerus, dan update sesuai kebutuhan. Maka ketika terdapat data baru, mesin akan langsung update weight dari tiap column sehingga adaptif terhadap perubahan terbaru, kelemahan dari pendekatan Online Learning adalah kemungkinan data yang masuk kotor / masih memiliki null value sehingga dapat mempengaruhi kualitas dari learning tersebut.

Menggunakan Model-based karena dari training data mesin dapat mengetahui model tiap data sehingga saat masuk data baru, dapat diprediksi model bentuk apa data baru tersebut dari model-model sebelumnya. Model-based lebih adaptif dibandingkan Instance-based karena parameter tidak di-tune oleh mesin saat terjadi perubahan pada Instance-based.

1. Link Github: <https://github.com/limasakeh/ML-Limas>