Night Login Open Membership 2022
Night Login Data Science & Data Analytics Community

Sefvia Lie
21/478184/TK/52695

1. Apa itu *Data Science, Data Analytics, Data Engineering, Business Analyst, dan Business Intelligence*? Sebutkan masing-masing perbedaannya!

   Data Science adalah bidang studi *advance* yang memerlukan pengetahuan dasar lain seperti statistik, *artificial intelligence*(AI), *programming skills*, *advanced analytics techniques* dan *scientific methods* untuk meng-ekstrak makna-makna penting dari sebuah data yang didapatkan dari website-website, *smartphones*, *customers*, *sensors* maupun sumber data-data lainnya untuk mempelajari pola maupun preferensi pengguna, yang nantinya dapat digunakan untuk membuat keputusan serta inovasi-inovasi baru.

   Data Analytics merupakan disiplin ilmu yang memiliki tujuan umum untuk mengekstrak makna-makna penting dari data, serta mengaplikasikan *statistical analysis and technologies* pada data untuk menemukan tren dan menyelesaikan masalah. *Data analysis* merupakan kumpulan proses, tools serta teknik untuk menganalisis data dan melakukan manajemen data, termasuk mengumpulkan data, mengorganisasi sebuah data dan menyimpan data.

   Data Engineering merupakan kumpulan proses untuk membuat sebuah *raw data* menjadi sebuah data yang dapat diolah dan dimengerti oleh para *data scientists* maupun grup dalam sebuah organisasi. Para *data engineers* memproses *raw data* serta membuat analisisnya untuk menghasilkan *predictive models*, yaitu model prediksi data di masa depan serta menampilkan tren data untuk jangka waktu yang singkat maupun panjang.

   Business Analyst merupakan sebuah pekerjaan yang menggunakan *data analytics* untuk menghubungkan *gap* antara IT dan bisnis yang sedang dijalankan, dengan tujuan untuk menilai sebuah proses, menentukan persyaratan, menyampaikan laporan serta rekomendasi untuk pengambilan keputusan berdasarkan data yang telah ada, untuk mencapai goals yang dimiliki oleh sebuah bisnis. *Business Analyst* meyakinkan para *business leaders* dan *user* bagaimana makna-makna penting yang didapatkan dari sebuah data dapat meningkatkan efisiensi serta dapat menambah nilai dari sebuah bisnis.

   Business Intelligence merupakan kumpulan prosedur dan teknik yang mengumpulkan, menyimpan serta menganalisis data yang dihasilkan dari kegiatan-kegiatan sebuah perusahaan. *Business Intelligence* mencakup penambangan data, proses analisis data, *benchmarking* kinerja sebuah bisnis serta analisis deskriptif dari data-data sebuah bisnis perusahaan.

Perbedaan antara Data Science, Data Analytics, Data Engineering, Business Analyst dan Business Intelligence adalah:
Data Science merupakan topik besar yang mencakup *data analytics*. *Data analytics* merupakan cabang dari *data science* yang berfokus pada menemukan jawaban spesifik dari pertanyaan-pertanyaan yang dihasilkan *data science*. Sedangkan data engineer bertugas untuk mengembangkan serta membuat design manajemen data dan memonitornya dalam sebuah perusahaan.
Perbedaan antara Business Analyst dan Business Intelligence adalah bahwa Business Intelligence berfokus pada analisis deskriptif, menjawab pertanyaan 'mengapa' dan 'bagaimana' serta menghasilkan kesimpulan berdasarkan data masa kini dan masa lampau untuk mengetahui hal yang sedang terjadi maupun yang telah terjadi. Sementara business analytics menggunakan data mining, modelling dan machine learning untuk menganalisa dan memprediksi kemungkinan yang akan terjadi di masa depan.

2. Sebutkan skill yang harus dimiliki oleh Data Scientist dan Data Analyst!
   Skills yang harus dimiliki oleh *data scientist*:
   - Dasar-dasar *data science*
   - Statistika
   - Menguasai *programming language* dan *programming knowledge*
   - Memanipulasi dan menganalisis data
   - Memvisualisasikan data
   - *Machine Learning*
   - *Deep Learning*
   - *Big Data*
   - *Software Engineering*
   - *Model Deployment*
   - *Interpersonal skills (Communication, storytelling skills)*
   - Pemikiran terstruktur
   - Rasa keingintahuan

   Skills yang harus dimiliki oleh data analyst:
   - Kemampuan matematis
   - Menguasai *programming language* seperti SQL, Oracle, Python
   - Kemampuan menganalisis, memodelkan dan menginterpretasi data
   - *Problem-solving skills*
   - *Approach* methodological dan logical
   - Kemampuan untuk mengorganisasikan jadwal serta *deadlines*
   - Accuracy dan ketelitian terhadap detail
   - *Interpersonal skills*
   - *Team Working skills*
   - *Written and verbal communication skills*

3. Sebutkan dan jelaskan jenis-jenis data berdasarkan sifatnya dan cara penyusunannya,
   dan berikan contohnya!
   Berdasarkan sifatnya, data dibedakan menjadi dua, yaitu data kualitatif dan data kuantitatif.

   Data kuantitatif, berdasarkan namanya merupakan data yang dapat diekspresikan dengan angka atau dapat dihitung dengan variabel numerik. Data kuantitatif dapat secara mudah dilakukan manipulasi statistik, serta dapat direpresentasikan dengan berbagai macam variasi grafik, seperti garis, bar, grafik, scatter plot, tabel, diagram, dan lainnya. Adapun contoh dari data kualitatif adalah temperatur tubuh manusia, nilai hasil ujian mahasiswa, ukuran sepatu konsumen, dan lainnya. Data kuantitatif juga terbagi menjadi dua kategori umum, yaitu discrete data yang merupakan data tetap,statis, tidak berubah-ubah serta continuous data, yang merupakan data yang dapat berubah-ubah seiring berjalannya waktu, atau yang dapat disebut data dinamis.

   Data kualitatif merupakan data yang tidak dapat diekspresikan dengan angka dan tidak dapat diukur. Data kualitatif terdiri dari huruf, kata, gambar, simbol dan lainnya, dan bukan angka. Karena tidak dapat dikategorikan berdasarkan angka, maka data kualitatif seringkali disebut sebagai categorical data. Data kualitatif biasanya menjawab pertanyaan bagaimana hal tersebut dapat terjadi dan mengapa hal tersebut dapat terjadi. Contoh dari data kualitatif adalah warna, destinasi liburan yang paling sering dikunjungi, warna kulit, dan lainnya.

   Berdasarkan cara penyusunannya, data dibedakan menjadi 4, yaitu data nominal, data ordinal, data interval dan data rasio.

   Data nominal merupakan data yang disusun, dikelompokkan atau dikategorikan berdasarkan pada peristiwa maupun waktu dimana data dikumpulkan. Data nominal digunakan untuk memberi label sebuah variabel. Contoh data nominal adalah gender, warna rambut, status, dan etnis.

   Data ordinal merupakan data yang diletakkan pada urutan tertentu berdasarkan posisinya atau berdasarkan sebuah skala. Data ordinal bisa saja menunjukkan peringkat atau kepentingan sesuatu. Contoh data ordinal adalah status ekonomi, peringkat dalam sebuah kompetisi, rating produk, dan lainnya.

   Data interval merupakan data yang menekankan pada urutan. Dalam pengurutannya, telah ditentukan interval atau perbandingan jarak antara satu data dan data lain. Contoh data interval adalah rata-rata berat badan wanita Indonesia adalah 50-55 kg.

   Data rasio merupakan data pengukuran yang memiliki nilai mutlak nol dalam perbandingan jaraknya sehingga data yang dimiliki akurat. Contoh data rasio adalah rasio pendaftar wanita dan pria adalah 1:1,5.

4. Sebutkan dan jelaskan jenis-jenis analitik dan berikan contohnya!
   Terdapat 4 jenis data analitik, yaitu Descriptive Analytics, Predictive Analytics, Prescriptive Analytics, dan Diagnostic Analytics.

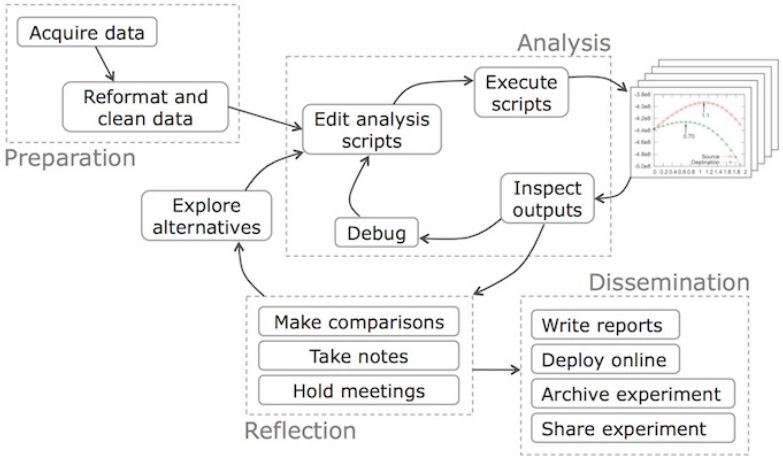
   Descriptive Analytics merupakan tipe analitik data yang lebih berfokus mencari jawaban dari pertanyaan apa sebuah peristiwa yang sedang terjadi, yang direpresentasikan dalam bentuk data. *Descriptive analytics* menganalisis data *real-time* dan data tentang peristiwa yang pernah terjadi sebelumnya untuk mengetahui tindakan apa yang harus dilakukan. Tujuan utama dari *descriptive analytics* adalah mencari alasan dibalik kesuksesan dan kegagalan dari masa lalu. *Descriptive analytics* adalah jenis analytics yang paling simple, yang memungkinkan data scientist untuk memecah data-data besar menjadi data yang lebih kecil dan lebih bermanfaat. Contoh dari *descriptive analytics* adalah data *monthly profit and loss statement* yang dimiliki oleh sebuah perusahaan.

   Predictive Analytics merupakan tipe analitik data yang berfokus untuk menjawab pertanyaan ‘Apa yang akan terjadi selanjutnya?’, memikirkan kemungkinan-kemungkinan yang akan terjadi di masa depan. Memiliki kemampuan untuk memprediksi data di masa depan memungkinkan kita untuk mengambil keputusan yang lebih baik. Contohnya adalah memprediksi dimana kemungkinan seseorang terkena serangan jantung meningkat ketika umurnya bertambah.

   Prescriptive Analytics merupakan tipe analitik data yang berfokus untuk menjawab pertanyaan ‘Apa yang harus dilakukan?’ setelah melihat data *predictive analytics*, untuk menentukan keputusan terbaik yang dapat diambil. Contoh dari *prescriptive analytics* adalah penggunaan untuk memilih rute tercepat dari sebuah perjalanan, menentukan jadwal ujian mahasiswa agar tidak bertabrakan dengan jadwal perkuliahan.

   Diagnostic Analytics merupakan tipe analitik data yang berfokus untuk mencari jawaban dari pertanyaan ‘Mengapa hal ini dapat terjadi?’. *Diagnostic analytics* membantu untuk menemukan anomali-anomali dan menemukan relasi dalam sebuah data, untuk mengetahui mengapa suatu peristiwa dapat terjadi. Contoh dari *diagnostic analytics* adalah menemukan pengaruh dari sebuah obat pada pasien yang memiliki keluhan yang beragam.

5. Sebutkan dan jelaskan alur yang dilakukan oleh Data Scientist atau Data Analyst untuk mencapai tujuan!



Untuk mencapai tujuan, seorang data scientist atau data analyst melakukan 4 tahapan, yaitu preparation, analysis, reflection dan dissemination.

Tahap pertama, yaitu *preparation phase,* terdiri dari step *acquire data* dan *reformat and clean data*. Sebelum melakukan analysis, *data scientist* harus terlebih dulu memperoleh (*acquire)* data tersebut dan melakukan *reformat* pada data tersebut menjadi format data yang dapat di komputasikan.

Selanjutnya dilakukan *analysis phase* yang terdiri dari *edit analysis scripts, execute scripts, inspect outputs, debugging dan re-editing*. Inti dari proses data science berada dalam tahap analisis ini, yaitu menulis, mengeksekusi dan memperbaiki, menyempurnakan program komputer, dengan melibatkan programming untuk menganalisis dan mendapatkan makna penting dari sebuah data.

Tahapan selanjutnya merupakan *reflection phase* yang terdiri dari *making comparisons, take notes* dan *hold meetings*. Fase ini melibatkan proses berpikir dan berkomunikasi mengenai output dari analisis yang telah dilakukan.

Tahapan terakhir merupakan tahapan *dissemination*, yang terdiri dari *writing reports, deploy online, archive experiment*, dan *sharing experiment. Data scientist* menulis laporan dalam bentuk memo internal, presentasi slideshow, *academic research*, atau bentuk laporan lainnya. Setelah itu, data scientist mendistribusikan software yang telah dibuat sehingga eksperimen yang telah dibuat dapat direproduksi atau dicoba *prototype systems*-nya oleh *data scientist* lain maupun masyarakat umum.

6. Apa perbedaan antara feature, variabel, dan kolom?

   Feature merupakan ekspresi dalam *machine learning* yang mengacu pada sebuah informasi ataupun data yang dapat diukur. Contohnya jika kita menyimpan usia, pendapatan, dan tinggi badan seseorang, dapat dikatakan bahwa kita menyimpan tiga '*feature*' dari orang tersebut.

   Sedangkan, dalam *data science*, *variable* merupakan simbol yang merepresentasikan *multiple data points*. Nilai sebuah variabel dapat

berubah-ubah pada data unit yang berbeda dalam sebuah populasi, dan nilai tersebut dapat berubah seiring berjalannya waktu. Contoh dari *variable* adalah pendapatan seseorang.

Sementara itu, dalam algoritma *machine learning*, dimana diasumsikan data yang kita miliki berupa sebuah tabel, kolom merepresentasikan feature atau atribut yang dimiliki oleh semua baris. *Column* merepresentasikan kategori dari sebuah informasi. Sedangkan baris merepresentasikan sebuah *event* atau *item* atau *instance*.

7. Apa itu Data Visualization dan Exploratory Data Analysis? Seberapa penting Data Visualization bagi seorang Data Scientist atau Data Analyst? Sebutkan jenis-jenis visualisasi dan gambarkan, serta jelaskan kegunaan visualisasi tersebut cocok untuk data yang sifat dan cara penyusunannya bagaimana?

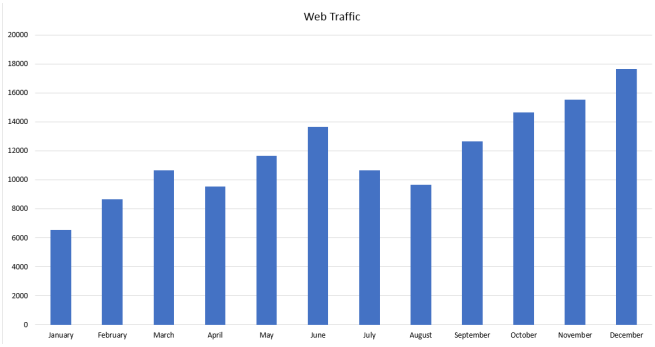
   Bagi seorang *data scientist* atau *data analyst*, *data visualization* merupakan serangkaian proses untuk menerjemahkan sebuah set data dan matriks agar menjadi representasi dari sebuah informasi ataupun data dalam bentuk sebuah grafik, diagram, gambar, ataupun bentuk representasi lainnya. *Data visualization* dapat memudahkan dalam proses identifikasi sebuah data, maupun memahami sebuah data sehingga akan lebih mudah untuk membagikan tren data yang sedang terjadi saat ini.

   Exploratory data analysis merupakan sebuah proses yang cukup kritikal untuk melakukan investigasi awal pada data untuk menemukan pola, anomali, mengetes hipotesis pada data serta mengecek asumsi dengan bantuan statistika dan representasi secara grafik, dengan tujuan utama untuk menemukan karakteristik utama dari sebuah data.

   Data Visualization merupakan proses yang sangat penting bagi seorang *data scientist* ataupun *data analyst*, karena *data visualization* dapat memudahkan dalam proses identifikasi sebuah data, memahami sebuah data dan membantu masyarakat umum seperti *business users* untuk memahami makna dari sebuah data. *Data visualization* dapat membantu untuk menemukan error dan *pattern* baru yang terdapat dalam sebuah data, yang nantinya akan memudahkan user untuk mengidentifikasi area yang perlu diperhatikan maupun dibenahi.

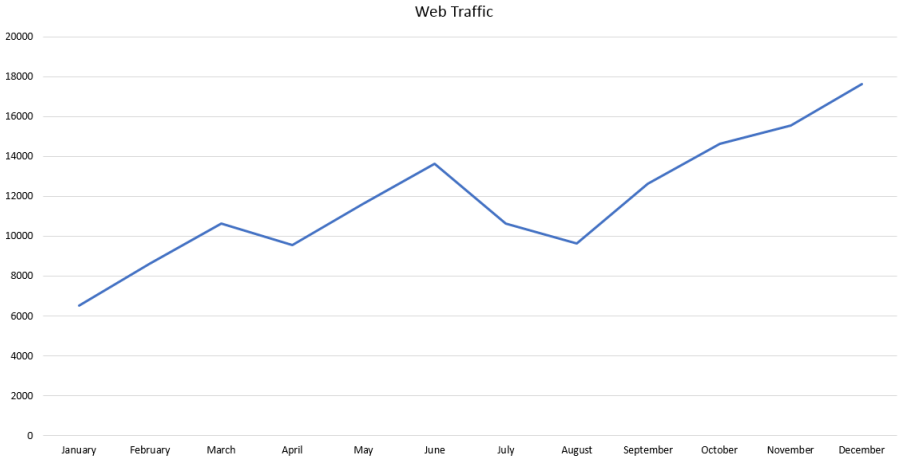
   Terdapat 5 jenis *data visualization* yang umum digunakan, yaitu *column charts, line graphs, matrix diagrams, scatter plot charts* dan *pie charts*.

a. Column Charts / Bar Charts / Graph



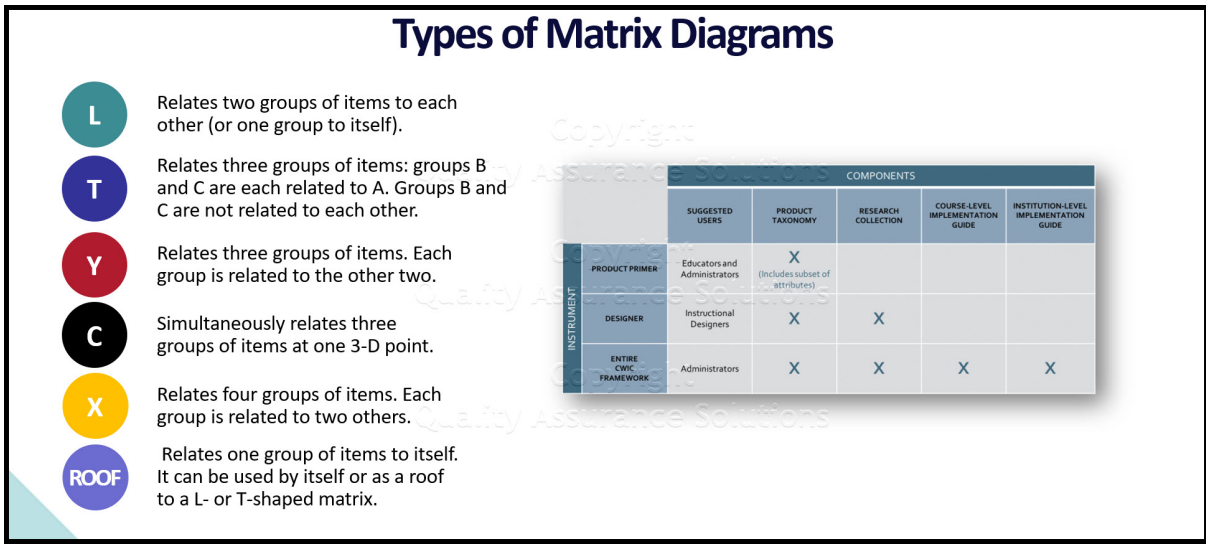
Tipe visualisasi *column charts* terdiri dari kolom-kolom horizontal maupun vertikal untuk merepresentasikan tiap set data. Tipe visualisasi ini cocok untuk menemukan perbedaan dari set-set data ataupun untuk mengidentifikasi bagaimana sebuah data berubah dari waktu ke waktu. Contoh penggunaan *column charts* adalah pada data pengunjung website per bulan, identifikasi audiens berdasarkan usia, serta mengidentifikasi penjualan product per bulannya. Cara penyusunannya adalah dengan memberikan label ataupun warna pada tiap set data untuk membedakannya. Dapat juga digunakan skala dalam bentuk angka.

b. Line Graphs



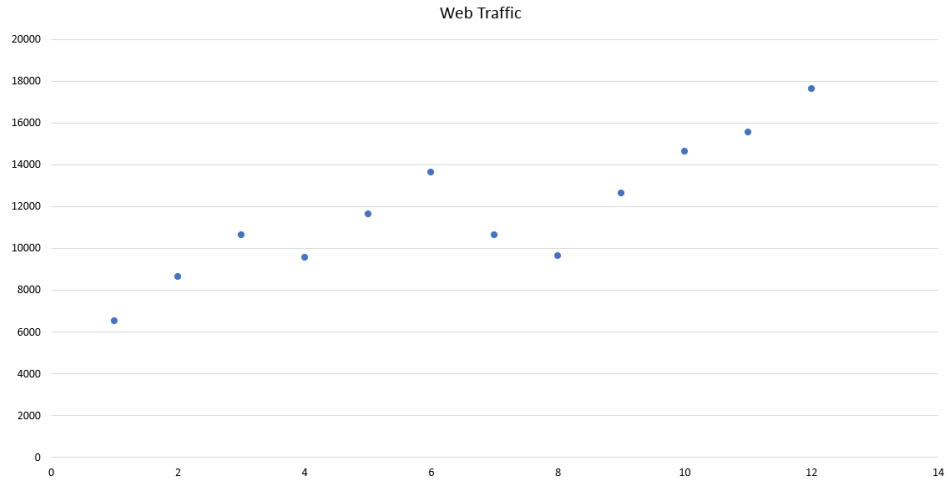
Tipe visualisasi *line graphs* merupakan tipe visualisasi yang umum digunakan untuk mengidentifikasi perubahan sebuah data dari waktu ke waktu. Tipe visualisasi ini cocok untuk data yang dinamis, continuous data atau data yang berubah-ubah seiring waktu. Cara penyusunannya dapat digunakan warna tiap garis yang berbeda untuk membedakan sebuah data. Pada umumnya, line graphs pada x-axis menunjukkan rentang waktu, sementara pada y-axis menunjukkan pengukuran yang sedang dilacak.

c. Matrix Diagrams



Tipe visualisasi *matrix diagrams* merupakan tipe visualisasi yang digunakan untuk mengidentifikasi hubungan antara data set yang berbeda. Matrix diagram menunjukkan perbedaan dari sebuah group data dalam kategori yang lebih besar, menunjukkan bagaimana sebuah grup data memiliki hubungan dengan group data lain. Cara penyusunannya berdasarkan grup data yang ingin dibandingkan. Terdapat 5 tipe matrix diagram, yaitu *L-shaped*, yaitu membandingkan dua grup data, atau membandingkan sebuah group data dengan dirinya sendiri. *Y-shaped*, yaitu membandingkan tiga group data yang terhubung satu sama lain. *C-shaped*, membandingkan tiga group data pada waktu yang sama. *T-shaped*, membandingkan dua group data yang terhubung satu sama lain. Dan *X-shaped*, membandingkan empat group data.

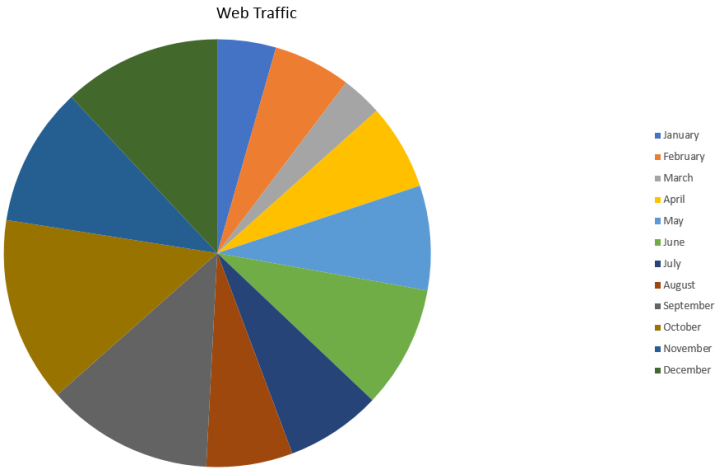
d. Scatter Plot Charts



Tipe visualisasi *scatter plot charts* merupakan tipe visualisasi untuk menemukan hubungan dari variabel-variabel dalam sebuah data menggunakan varied data points. Tipe visualisasi ini cocok digunakan pada data yang memiliki variabel-variabel yang terhubung satu sama

lain. Penyusunannya adalah dengan melakukan plotting data, mirip seperti fungsi dalam matematika. Y = f(x). Contoh penggunaannya adalah menemukan apakah jam kuliah menentukan tingkat pemahaman mahasiswa, berdasarkan hasil ujian yang didapatkan.

e. Pie Charts



Tipe visualisasi *pie charts* seringkali digunakan pada tipe data yang memiliki persentase yang menunjukkan dominansi. Tipe visualisasi ini cocok digunakan untuk tipe data yang dapat direpresentasikan dalam bentuk persentase (*percentage)*. Cara penyusunannya adalah dengan memberi label pada setiap data, memastikan persentase tiap data mencapai 100% lalu membuat *pie chart*nya, agar audiens dapat memahami data mana yang lebih mendominasi. Contoh penggunaanya adalah mengidentifikasi bulan dimana pengunjung website mencapai maksimum.

Selain kelima tipe visualisasi data tersebut, terdapat pula tipe visualisasi data berdasarkan area dan ukurannya, visualisasi data berdasarkan warna seperti *heatmap* pada peta, visualisasi data berdasarkan gambar, visualisasi data berdasarkan konsep serta visualisasi data berdasarkan grafik atau bagan.

8. Jelaskan apa itu *Artificial Intelligence* (AI)! Jelaskan pula apa kaitannya dengan *Data Science* dan *Data Analytics*!
   Artificial Intelligence adalah sebuah teori dan pengembangan sistem komputer agar dapat menyelesaikan permasalahan-permasalahan umumnya yang membutuhkan kecerdasan manusia, seperti *visual perception, speech recognition*, proses pengambilan keputusan dan menerjemahkan bahasa. *Artificial Intelligence* yang telah dapat kita temui adalah seperti rekomendasi video *youtube, siri, alexa, google assistant, self-driving car* seperti tesla, *email spam filters*, dan lainnya.
   Kaitan antara *Artificial Intelligence* dan *Data Science* adalah bahwa *Data Science* melakukan proses untuk mengekstrak makna maupun pengetahuan

yang dapat didapatkan dari data, untuk membantu *Artificial Intelligence* mencari jawaban dari permasalahan-permasalahan yang membutuhkan kecerdasan manusia, berdasarkan pola-pola yang telah didapatkan dari data. Kaitan antara *Artificial Intelligence* dan *Data Analytics* adalah *data analytics* merupakan teknologi untuk mempelajari data dan pola-pola yang terdapat dalam data untuk menyelesaikan permasalahan-permasalahan dalam *artificial intelligence*.

9. Apa itu *machine learning*? Sebutkan dan jelaskan tipe dari *machine learning* beserta contoh algoritmanya!

   Machine learning merupakan salah satu cabang dari *Artificial Intelligence* dan *computer science* yang berfokus pada penggunaan data dan algoritma untuk meniru bagaimana manusia belajar dan beradaptasi tanpa instruksi secara eksplisit, dengan menggunakan algoritma dan model statistik untuk menganalisis dan menarik kesimpulan dari pola-pola yang terdapat di data, dengan tujuan utama untuk meningkatkan akurasi.
   Terdapat empat tipe dari *machine learning*, yaitu *supervised learning, semi-supervised learning, unsupervised learning* serta *reinforcement learning*.

   Supervised learning merupakan tipe *machine learning* yang bertujuan untuk memberikan label pada dataset untuk melatih algoritma, mengklasifikasi data ataupun memprediksi hasil secara akurat, berdasarkan data historis. Contoh algoritma supervised learning adalah *linear regression* untuk permasalahan regresi, *neural networks, naive bayes, logistic regression, random forest, support vector machine*, dan lainnya.

   Unsupervised learning merupakan merupakan tipe *machine learning* yang menggunakan algoritma untuk menganalisis dan mengelompokkan dataset yang belum memiliki label. Algoritma tersebut memiliki tujuan untuk mengidentifikasi pola-pola tersembunyi dalam sebuah data maupun melakukan pengelompokkan data tanpa bantuan manusia. Contoh algoritma *unsupervised learning* adalah *exploratory data analysis, cross-selling strategies, customer segmentation,* serta *image and pattern recognition*.

   Semi-supervised learning merupakan tipe *machine learning* yang berada di tengah-tengah antara *supervised learning* dan *unsupervised learning*. *Semi-supervised learning* mencari solusi dari masalah data yang tidak mempunyai label yang cukup jelas, atau untuk melatih *supervised learning algorithm*. Contoh algoritma *semi-supervised learning* adalah *text document classifier*.

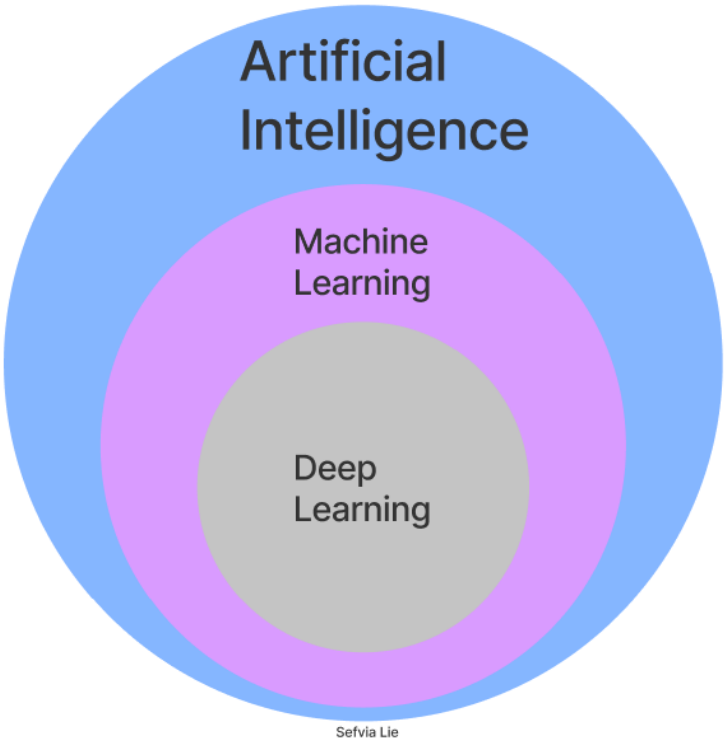
   Reinforcement learning merupakan tipe *machine learning* yang biasanya digunakan untuk robotik, pembuatan game dan navigasi. Tipe *machine learning* ini memungkinkan algoritma untuk menemukan *action* yang akan menghasilkan output terbaik dari hasil *trial and error*. Contoh algoritma *reinforcement learning* adalah *decision tree*.

10. Jelaskan apa itu *Deep Learning*!

    Deep Learning merupakan sebuah subset dari *machine learning*, yang

merupakan *neural network* dengan tiga layer. *Neural networks* ini mencoba untuk memanipulasi, melakukan simulasi, menirukan perilaku otak manusia, yaitu memahami dan mempelajari makna dari sebuah data. *Deep learning* mampu melakukan aproksimasi prediksi, serta mengoptimalkan keakuratan prediksi tersebut. *Deep learning* merupakan aplikasi dari *artificial intelligence* (AI) yang meningkatkan otomatisasi, melakukan tugas serta memecahkan suatu masalah tanpa melibatkan manusia. Banyak produk yang sering kita temui sehari-hari yang menerapkan *deep learning*, contohnya adalah *digital assistant* pada handphone maupun gadget kita (*siri, alexa, bixby, google assistant*), *self driving car* seperti tesla, dan lainnya.

11. Buat visualisasi yang menunjukan kedudukan AI, *Machine Learning*, dan *Deep Learning*!



12. Jika Anda sudah menjawab pertanyaan di atas, seharusnya Anda sudah dapat memahami istilah-istilah di bawah ini, jika belum, coba cari pada referensi terpercaya, kemudian pahami dan catat, jika sudah, centang pada istilah-istilah yang sudah dipahami.

☑ ~~● Data Preprocessing~~
☑ ~~● Data Transformation~~
☑ ~~● Data Cleaning~~
☑ ~~● Data Latih~~
☑ ~~● Data Tes~~

- ☑ ~~● Missing Value~~
- ☑ ~~● Outlier~~
- ☑ ~~● Normalization~~
- ☑ ~~● Label Encoding~~
- ☑ ~~● One Hot Encoding~~
- ☑ ~~● Exploratory Data Analysis~~
- ☑ ~~● Feature Engineering~~
- ☑ ~~● Feature Selection~~
- ☑ ~~● Koefisien korelasi~~
- ☑ ~~● Modelling~~
- ☑ ~~● Regresi~~
- ☑ ~~● Klasifikasi~~
- ☑ ~~● Klasterisasi~~
- ☑ ~~● Cross Validation~~
- ☑ ~~● Presisi, recall, akurasi, F1 Score~~
- ☑ ~~● RMSE, MAE, MAPE~~

Tools and Environment

1. Apa itu bahasa pemrograman Python?

Python merupakan salah satu bahasa pemrograman dari banyaknya bahasa pemrograman yang ada saat ini. Bahasa pemrograman ini pertama kali diperkenalkan pada 20 Februari 1991, yaitu 30 tahun lampau. Dibandingkan bahasa pemrograman lain seperti C, C++ dan C#, bahasa pemrograman python dapat dibilang cukup terkini. Bahasa pemrograman ini terbilang bahasa pemrograman yang cukup mudah dipelajari, sehingga digemari sebagian besar kalangan masyarakat.

Berdasarkan website resminya, bahasa pemrograman python merupakan bahasa pemrograman yang interpreted, berorientasi pada objek, bahasa pemrograman tingkat tinggi yang dinamis, dapat digunakan seiring berjalannya waktu. Bahasa pemrograman ini banyak digunakan untuk mengembangkan aplikasi, maupun bahasa yang dapat menghubungkan berbagai komponen. Bahasa pemrograman ini merupakan bahasa pemrograman yang mengutamakan readability dari codenya, agar dapat dipahami oleh orang awam sekalipun. Bahasa pemrograman yang berorientasi objek ini memiliki tujuan untuk membantu programmer untuk menulis kode yang jelas, yang dapat dibaca oleh masyarakat umum untuk proyek skala kecil dan besar sekalipun.

2. Apa itu Jupyter Notebook?

Jupyter notebook merupakan sebuah open source web application yang dapat digunakan untuk membuat dan membagikan dokumen-dokumen yang berisi code terutama dalam bahasa pemrograman Python, persamaan matematika, visualisasi, serta text yang digabung dalam satu file interaktif. Notebook buatan

Jupyter tersebut berisi dokumen yang dapat dibaca oleh manusia yang mengandung analisis serta visualisasinya, serta berisi executable documents yang dapat di run untuk melakukan analisis data. Notebook ini dapat di-run oleh siapapun yang membukanya, dapat pula dilakukan reproduksi eksekusi code di dalamnya.

3. Apa itu Kaggle?

Kaggle merupakan sebuah website berisi komunitas online dari data scientist, developers serta machine learning practitioners. Dalam website ini, para data scientist dapat menemukan dan menggunakan lebih dari 50.000 public datasets yang tersedia, serta lebih dari 400.000 public notebooks yang dapat digunakan sebagai referensi dalam melakukan analisis data. Kaggle juga merupakan sebuah situs pencarian seperti google, tempat para data scientist menemukan, menerbitkan data sets yang mereka miliki, serta melakukan eksplorasi dan membuat model data pada web-based data science-environment. Website ini juga memungkinkan beberapa data scientist untuk melakukan kolaborasi dengan data scientist lain maupun machine learning engineers lain. Terdapat pula kompetisi-kompetisi yang dapat diikuti oleh pengguna untuk menyelesaikan permasalahan yang terkait dengan data science.

4. Apa itu Google Colab?

Google Colab merupakan sebuah produk dari *google research*, yang memungkinkan para pengguna untuk menulis dan melakukan *run code* dalam bahasa pemrograman Python menggunakan browser, yang ditujukan secara khusus untuk *machine learning, data analysis* dan *education*. Colab juga merupakan produk yang menggunakan *Jupyter* sebagai *base* dari *open source*-nya. Colab memungkinkan user untuk memakai dan membagikan Jupyter notebook mereka kepada orang lain tanpa mengharuskan pengguna untuk mendownload, menginstall ataupun melakukan run. Colab notebooks disimpan pada Google Drive pengguna, juga dapat pula dibuka melalui akun GitHub. Google Colab merupakan produk yang memberikan *free access* kepada pengguna untuk melakukan komputasi pada *resource-resource*, termasuk GPU.

5.Sebutkan library dasar yang digunakan dalam bahasa pemrograman Python untuk menyelesaikan tugas sebagai seorang *Data Scientist dan Data Analyst*!

Library dasar dalam bahasa pemrograman Python yang sering digunakan oleh seorang *Data Scientist*:

1. NumPy
2. Theano
3. Keras
4. PyTorch
5. SciPy
6. PANDAS
7. PyBrain
8. SciKit-Learn

9. Matplotlib
10. Tensorflow
11. Seaborn
12. Bokeh
13. Plotly
14. NLTK (Natural Language ToolKit)
15. Gensim
16. Scrapy
17. Statsmodels
18. Kivy
19. PyQt
20. OpenCV

Library dasar dalam bahasa pemrograman Python yang sering digunakan oleh seorang Data Analyst:

1. Numpy
2. Spicy
3. Pandas
4. Matplotlib
5. Scikit
6. StatsModels
7. Seaborn

Case Study

Anda sudah mempelajari pengetahuan dasar tentang Data Science dan Data Analytics, serta mempelajari tools dan environment yang diperlukan bagi seorang Data Scientist dan Data Analytics. Sekarang Anda diminta untuk mengamati kasus yang diselesaikan dengan Data Science dan Data Analytics. Buka https://www.kaggle.com/c/titanic/code pilih salah satu kemudian amati, pahami, dan rangkum sesuai dengan apa-apa yang Anda pelajari pada bagian sebelumnya, tidak ada format khusus dalam rangkuman pada bagian ini. Jika Anda ingin melakukan proses analitik dari 0, itu lebih baik, namun jika tidak memungkinkan silahkan amati, pahami, kemudian rangkum sesuai dengan apa-apa yang Anda pelajari pada bagian sebelumnya.

Berdasarkan website, saya mempelajari bagaimana melakukan *import library* Python yang diperlukan seperti *numpy* dan *pandas* sebagai *data analysis libraries* dan *matplotlib* dan *seaborn* sebagai *visualization libraries*. Selain itu, saya mempelajari cara melakukan *ignore* pada *warnings*.
Selanjutnya, saya mempelajari bagaimana membaca training and testing data yang berada dalam format CSV. Saya mempelajari *function* describe() menggunakan pd.read_csv. Saya juga mempelajari *feature*s yang terdapat dalam dataset Titanic serta cara melengkapinya. *Feature* pada data dibagi menjadi tiga kategori penting, yaitu:

1. *Numerical Features*: Age(Continuous)[float], Fare(Continuous), SibSp(Discrete)[int], Parch(Discrete)[int]
2. *Categorical Features*: Survived [int] , Sex[string], Embarked[string], Pclass [int]
3. *Alphanumeric Features*: Ticket[string], Cabin[string]

Summary / Kesimpulan dari data set Titanic:
1. Terdapat 891 penumpang
2. *Feature* usia yang hilang adalah sebanyak 19.8% dari *value*nya. Karena usia merupakan salah satu faktor penting dari kemungkinan selamatnya seseorang, kita harus mengisi data usia yang hilang.
3. *Feature cabin* yang hilang adalah sebanyak 77.1% dari *value* aslinya. Karena terlalu banyak data yang hilang, akan sulit apabila kita mengaproksimasi dan mengisi data tersebut, maka kita tidak akan memasukkan data *Cabin* dalam *dataset*.
4. Feature *embarked* yang hilang adalah sebanyak 0.22% dari *value* aslinya, sehingga tidak akan membawa dampak yang begitu besar.

Prediksi berdasarkan data yang telah diperoleh adalah:
1. Berdasarkan jenis kelamin: Wanita memiliki kemungkinan lebih besar untuk selamat.
2. Berdasarkan *SibSp/Parch*: Orang yang travelling sendirian memiliki kemungkinan yang lebih besar untuk selamat, karena hanya mementingkan dirinya sendiri.
3. Berdasarkan Usia: Orang dengan usia yang lebih muda, seperti anak kecil memiliki kemungkinan yang lebih besar untuk selamat karena didahulukan oleh para orang tua.
4. Berdasarkan *Pclass*: Orang dengan status sosial / kelas sosial yang lebih tinggi memiliki kemungkinan yang lebih besar untuk selamat.

Selanjutnya dilakukan *data visualization* berdasarkan data yang telah dimiliki untuk melihat apakah prediksi yang telah dibuat akurat, dengan menggunakan bahasa pemrograman Python. Terbukti bahwa persentase wanita yang selamat mencapai 74%, sementara laki-laki hanya 18%.
Terbukti pula bahwa orang dengan status sosial yang lebih tinggi memiliki persentase 62% untuk selamat. Orang yang mempunyai banyak kerabat di dalam kapal memiliki kemungkinan yang lebih kecil untuk selamat, serta anak kecil maupun anak bayi memiliki kemungkinan paling tinggi untuk selamat.
Selain itu, ditemukan bahwa orang dengan nomor *Cabin* yang terekam di data memiliki kemungkinan yang lebih besar untuk selamat. *Data visualization* ini menggunakan tipe *bar plot*, dengan bahasa pemrograman Python.

Setelah itu, saya mempelajari cara melakukan *cleaning data* untuk menemukan data yang hilang serta menghapus data yang tidak penting. Dari proses *cleaning data*, ditemukan fakta bahwa:

1. Terdapat total 418 penumpang

2. 1 *value* dalam feature *Fare* menghilang
3. Sekitar 20.5% data pada *feature* Usia yang harus kita dapatkan, menghilang.
4. Mayoritas penumpang menaiki kapal dari *Southampton* (S)

Selain itu dilakukan pengisian data pada age feature yang hilang dengan melakukan prediksi berdasarkan *Title* yang dimiliki (Mr, Miss, Mrs, Baby), *name feature, sex feature, embarked feature*, serta *test feature*.

Selanjutnya dilakukan pemilihan untuk menentukan model mana yang paling cocok dan paling baik. Dilakukan splitting training data sebesar 22% untuk menguji akurasi dari model-model yang berbeda.
Dilakukan testing model:
1. Gaussian Naive Bayes
2. Logistic Regression
3. Support Vector Machines
4. Perceptron
5. Decision Tree Classifier
6. Random Forest Classifier
7. KNN or k-Nearest Neighbors
8. Stochastic Gradient Descent
9. Gradient Boosting Classifier

Untuk tiap model, digunakan 80% training data dan dilakukan prediksi pada 20% data sisanya untuk mengecek akurasinya.

Ditemukan bahwa model yang memiliki tingkat akurasi tertinggi adalah *Gradient Boosting Classifier* dengan skor 84,77. Sehingga diambil keputusan untuk menggunakan model *Gradient Boosting Classifier* untuk *testing data*.

Tahap terakhir dari proses ini adalah membuat *Submission file* dalam format csv, yaitu submission.csv.