zenius × Kampus Merdeka INDONESIA JAYA

Getting Started in Data Science

Day 2

Machine Learning Class

Program Studi Independen Bersertifikat
Zenius Bersama Kampus Merdeka

Data Analysts does not need to learn statistical modelling, because that is for Data Scientists

A. Benar

B. Salah

Data Analysts does not need to learn statistical modelling, because that is for Data Scientists

A. Benar

B. Salah

While more complex modelling job is mainly done by Data Scientists, Data Analysts should also have a strong basic foundation in statistics.

Data Scientist who does Machine Learning does not need to learn SQL

A. Benar

B. Salah

Data Scientist who does Machine Learning does not need to learn SQL

A. Benar

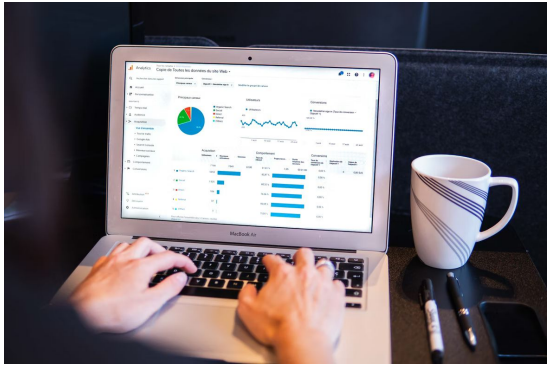
B. Salah

In tech companies with very big volume of data, getting training data for your ML is done by querying your companies' RDBMS

1. 3 Types of Data Talents
2. How Data Team Work Together in a Company
3. How to Create a Data Science Portfolio
4. Github for Hosting Portfolio
5. Getting Data from Kaggle
6. Building Self-Learning Habits

3 Types of Data Talents

3 Types of Data Talents



Data Analyst: Dashboarding, Querying, Analysis

Stack: SQL, Tableau, Power BI

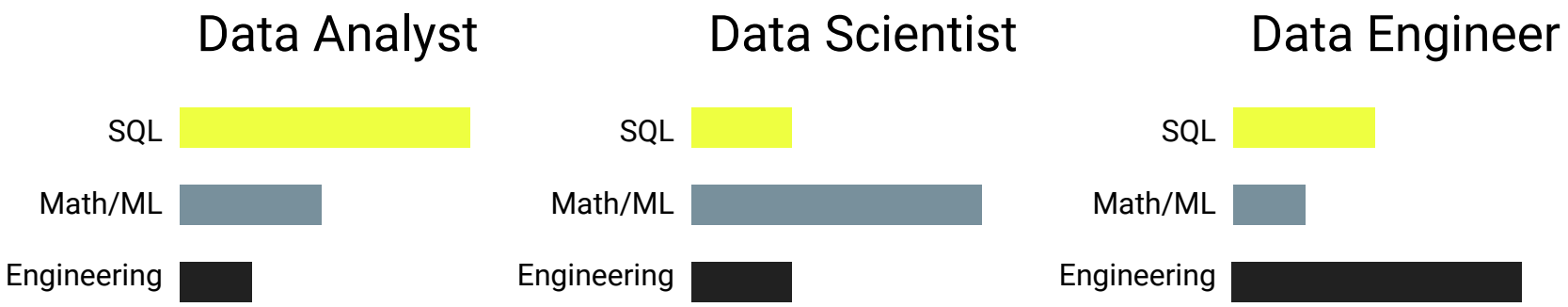
Data Scientist: Mathematical Modelling, Machine Learning

Stack: Python, SQL

Data Engineer: Maintain Cloud & Data Infrastructure and Pipelines

Stack: SQL, Hadoop, Spark

3 Types of Data Talents



How Data Team Work Together in a Company?

DS Team in Netflix: An Example

1. Data Analyst: Gather Customer Behavior, Visualize Types of Movies, Create Dashboard, etc
2. Data Scientist: Creating a Recommendation Engine, or a 'Movie Predictor'
3. Data Engineer: Manages how data is stored, and how model will later be deployed

Data Science Portfolio

Why do we need portfolio?

Importance of Portfolio:

- Shows that you know Data Science even though you haven't had a Data Science job
- Showcases technical and presentation skills, which are important as a Data Scientist

How does a good portfolio look like?

1. Contains an end-to-end project. A project that starts with importing data, cleaning the data, EDA (Exploratory Data Analysis), then to modelling, evaluating the model, and having a good, well-written summary.
2. Published in at least github. Will be much better if you write about it in LinkedIn, or in a Towards Data Science article. Or create your own website.

How does a good portfolio look like?

Examples:

- https://github.com/Radvian/apple-img
- https://harrisonjansma.com/

How does a good portfolio look like?

Guidelines:

1. Choose your topic.
2. Choose your data. (Or collect your own data, then host it somewhere so people can see it too)
3. Have a vision, what do you want to create with your data?
4. Create it!
5. Upload it to github, and write about it.

Github for Hosting Portfolio

What is Github?

Github is:

- A platform to host your codes.
- For companies - this will be where they store their codes
- For individuals - this will be where you upload your personal projects!

What is Github?

We are not here to discuss about using github to collaborate in a company - we are here to discuss about using github to host your Data Science projects.

What is Github?

Hands On:

Creating your first repository and uploading your files into github!

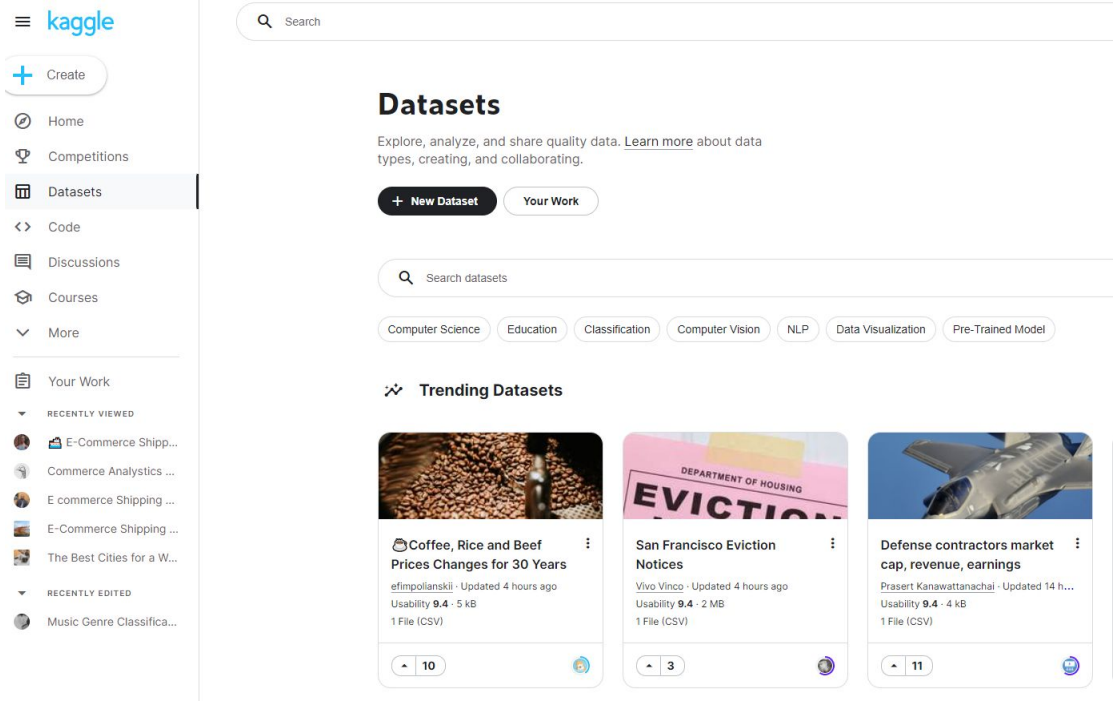
Getting Data from Kaggle

What is Kaggle?

A place for Data Scientists to:

- Get data from various institutions
- Participate in Competitions
- Upload and share their data to the world!

What is Kaggle?



Bottom Down Approach

We have a collection of transactions data. We initially don't know what to make use of it.

After doing exploratory data analysis, we notice a suspicious pattern.

We bring the data to the security / payment team, and they think it's fraudulent.

We then gather more data, add features, and try to come up with a Machine Learning model to detect frauds.

Kaggle Alternatives

- https://archive.ics.uci.edu/
- https://www.google.com/publicdata/directory
- https://datasetsearch.research.google.com/

Building Self Learning Habits

Self Learning Habits

Most important thing to have as a Data Scientist.

Why?

After this class ends, after you graduate from university, you will not have ‘teacher’ to teach you again.

But…

The data science world continues to advance and evolve - it won’t wait for you.

Self Learning Habits

If you only can learn while you're inside a bootcamp / university, then forget about being a Data Scientist.

Data Science is a constantly improving field - with new frameworks, machine learning models, papers, research, being discovered day by day.

Furthermore, coding work requires you to often troubleshoot for errors.

How to develop this habit?

You just need a correct mindset:

1. Don't be afraid of failures or encountering error. You will make mistakes, period.
2. Be sure that everything you want to know, most likely exists somewhere in the internet - you just need to Google it.
3. Find your own method to remember what you've learned.

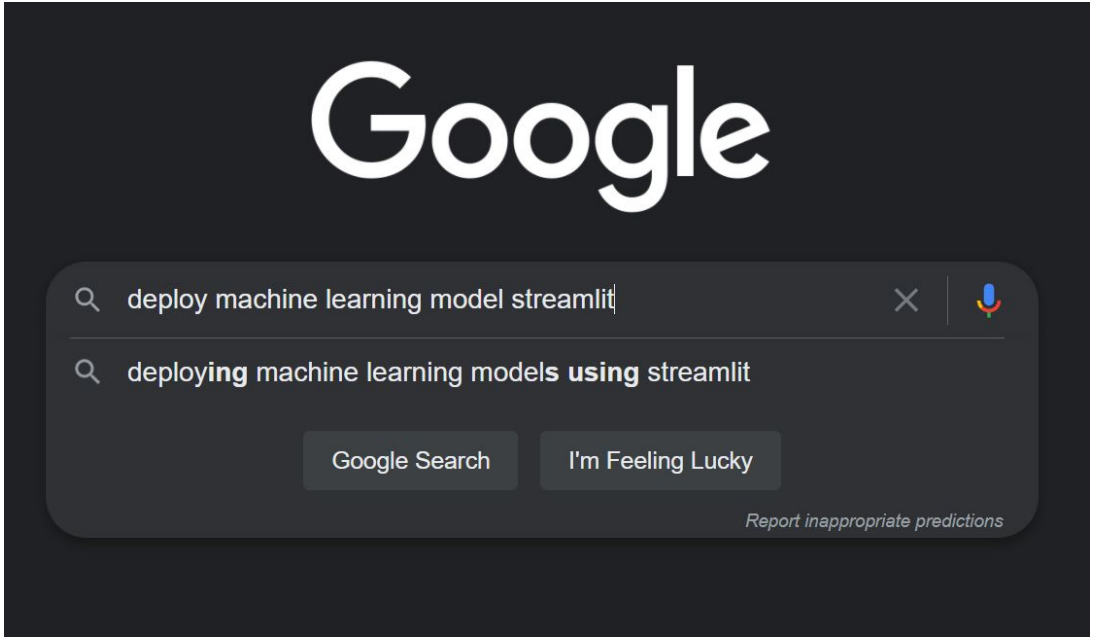
Example

You want to know how to deploy your Machine Learning project like my portfolio project.

However, this is not taught in this bootcamp - nor in your university.

Example

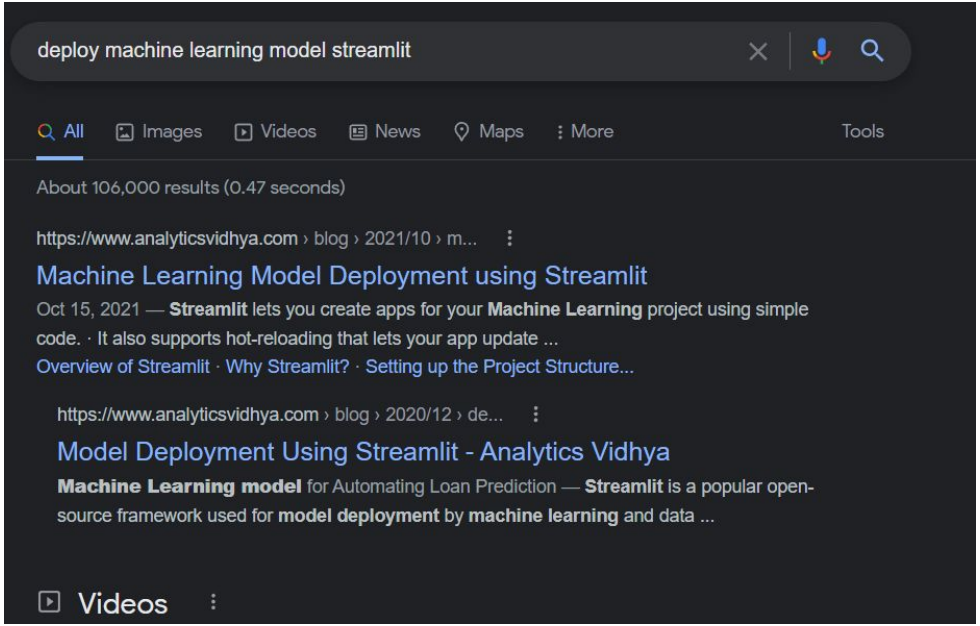
Step 1: “Just Google it!”



Example

Step 2:

Read 1-3 top articles that you found. The more you like to Google stuffs, the more you get a sense on which websites to read.

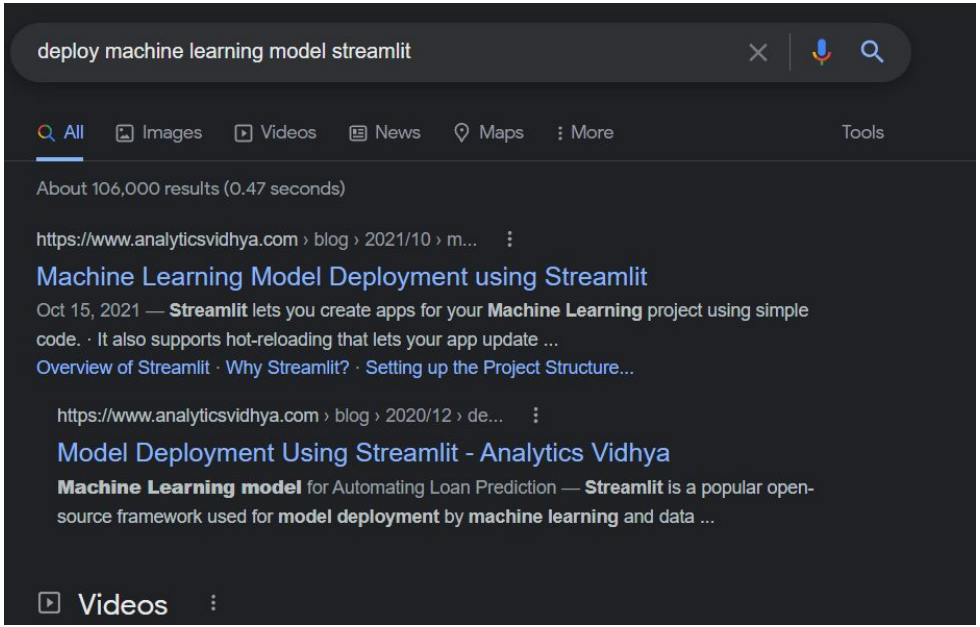


Example

Step 2 (cont):

For Data Science, usually good beginner tutorials for everything can be found in websites:

- Analytics Vidhya
- Towards Data Science
- Machine Learning Mastery



Example

Step 3:

- Usually in these articles, there are codes.
- Just follow (copy-paste) these codes and run it in your machine!
- Then, learn what does each line of code means, and try to understand how things are interconnected.

Example

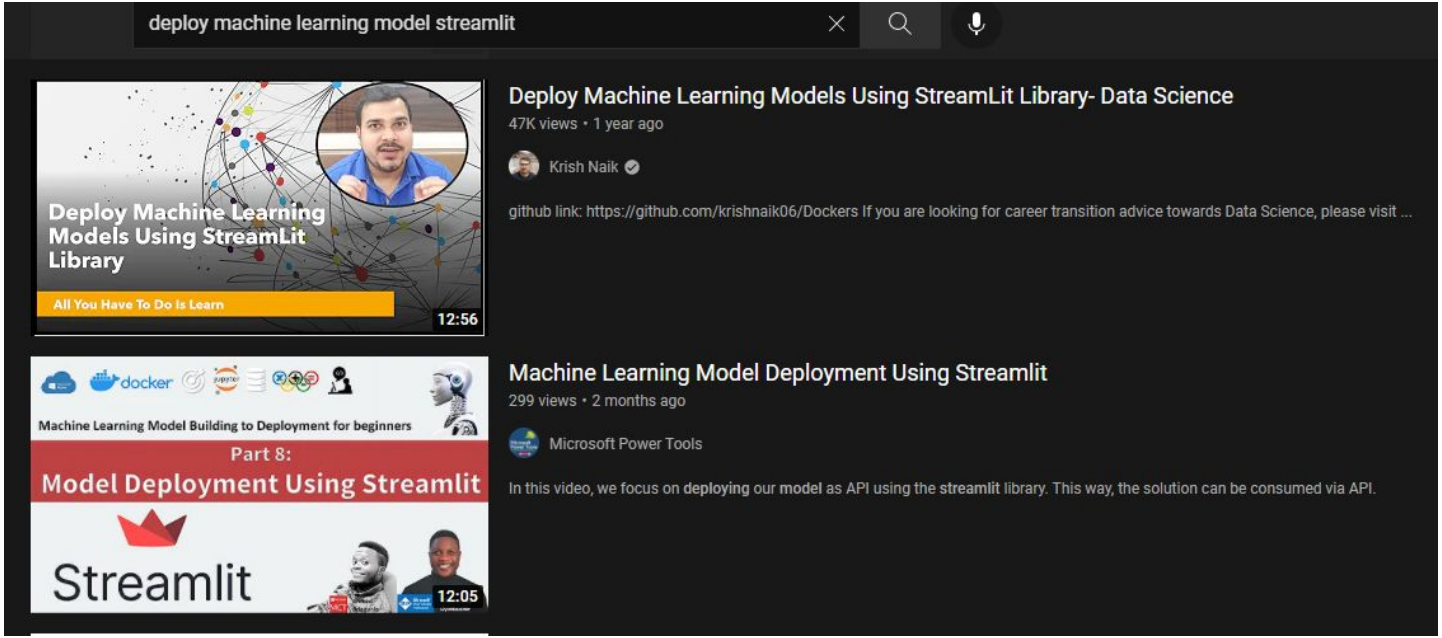
Step 4:

If you encounter error, you could search your errors in: https://stackoverflow.com/

If you learn better by watching a video rather than reading, then search it on YouTube!

Yes! YouTube isn't only to watch podcasts or music videos - there are A LOT of Data Science tutorials in YouTube that you can just...follow along.

Example



Summary

Summary

- Data Talents will be increasingly needed as businesses are in the process of digitization.
- You need portfolio containing solid projects to increase your chances of getting recruited.
- Having a self-learning habit is a very important aspect of a successful future data scientists.

Terima kasih!

Ada pertanyaan?