Final Project

12S3205 - Penambangan Data

Semester Genap 2024/2025

Overview

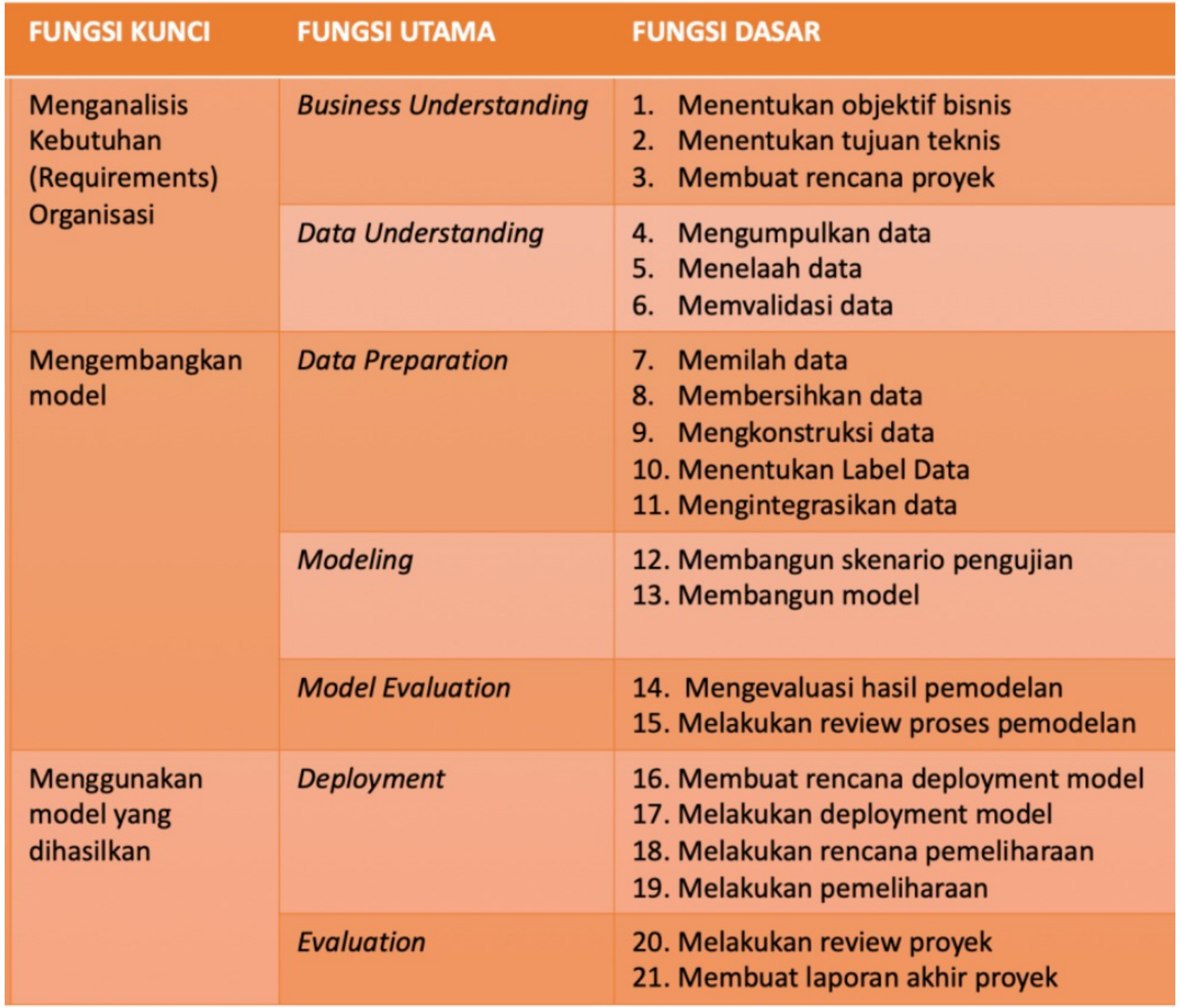
1. Task
2. Steps
3. Cases
4. Timeline
5. Submission

Tasks

1. Form a group of 3-4 students.
2. Build data mining project with CRISP-DM methodologi (Standard Kompetensi Kerja Nasional: Keputusan Menteri Ketenagakerjaan No 299 thn 2020)
3. Report your work (deliverables).

Steps:

Standard Kompetensi Kerja Nasional: Keputusan Menteri Ketenagakerjaan No 299 thn 2020



FUNGSI KUNCI	FUNGSI UTAMA	FUNGSI DASAR
Menganalisis Kebutuhan (Requirements) Organisasi	*Business Understanding*	1. Menentukan objektif bisnis 2. Menentukan tujuan teknis 3. Membuat rencana proyek
	Data Understanding	4. Mengumpulkan data 5. Menelaah data 6. Memvalidasi data
Mengembangkan model	*Data Preparation*	7. Memilah data 8. Membersihkan data 9. Mengkonstruksi data 10. Menentukan Label Data 11. Mengintegrasikan data
	Modeling	12. Membangun skenario pengujian 13. Membangun model
	Model Evaluation	14. Mengevaluasi hasil pemodelan 15. Melakukan review proses pemodelan
Menggunakan model yang dihasilkan	*Deployment*	16. Membuat rencana deployment model 17. Melakukan deployment model 18. Melakukan rencana pemeliharaan 19. Melakukan pemeliharaan
	Evaluation	20. Melakukan review proyek 21. Membuat laporan akhir proyek

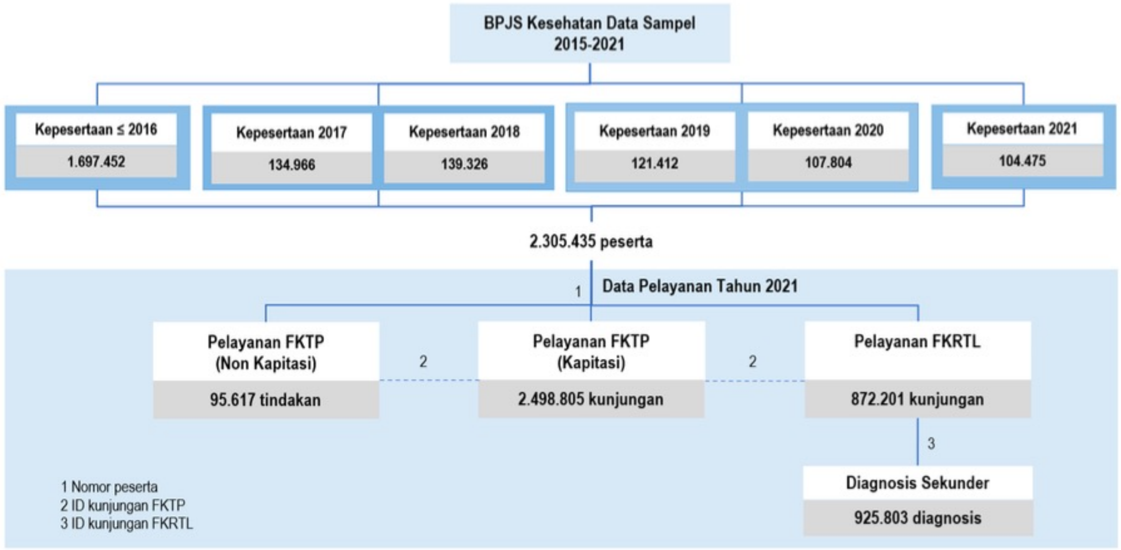
Case: BPPJS Hackaton

- Case 1: Sample BPJS 2015 – 2021
- Case 2: Exploring Mental Health Data
- Case 3: House Prices - Advanced Regression Techniques
- Case 4: Location-based species presence prediction
- Case 5: Stanford RNA 3D Folding

Case 1: BPPJS Hackaton

- *Case Sample BPJS 2015 – 2021*
  - *Supervised, Semi-Supervised dan Unsupervised Learning*
  - *Mengembangkan sebuah model data mining untuk melakukan klasifikasi, prediksi, maupun klastering pada sample BPJS data 2015-2021*
- *data:*

https://drive.google.com/drive/folders/1eNDeoaPfQJQWr5RUSVdgnobSXQ5VsKRU?usp=sharing



BPJS Kesehatan Data Sampel 2015-2021

Kepesertaan ≤ 2016 1.697.452

Kepesertaan 2017 134.966

Kepesertaan 2018 139.326

Kepesertaan 2019 121.412

Kepesertaan 2020 107.804

Kepesertaan 2021 104.475

2.305.435 peserta

1 Data Pelayanan Tahun 2021

Pelayanan FKTP (Non Kapitasi) 95.617 tindakan

2

Pelayanan FKTP (Kapitasi) 2.498.805 kunjungan

2

Pelayanan FKRTL 872.201 kunjungan

3

Diagnosis Sekunder 925.803 diagnosis

1 Nomor peserta
2 ID kunjungan FKTP
3 ID kunjungan FKRTL

Gambar 4 Hirarki data sampel BPJS Kesehatan 2015-2021

Case 1: BPPJS Hackaton (Requirement)

Case 1:

- Untuk kategori ketiga, silahkan anda bagi menjadi data training dan data validation berbeda dan memenuhi evaluasi berikut:
  - Classification problem: Precision > 0.60, Accuracy > 0.60, Recall > 0.65
  - Regression problem: MAE < 900 dan MAPE < 70%
  - Clustering problem: SC > 55%
- Anda diijinkan untuk menggunakan data lain yang diunduh secara legal untuk melakukan kombinasi atau menambah data dengan data sample yang diberikan

Case 2: Exploring Mental Health Data

- *https://www.kaggle.com/competitions/playground-series-s4e11/overview*
- Goal: to use data from a mental health survey to explore factors that may cause individuals to experience depression.
- **About the Tabular Playground Series**

The goal of the Tabular Playground Series is to provide the Kaggle community with a variety of fairly light-weight challenges that can be used to learn and sharpen skills in different aspects of machine learning and data science. The duration of each competition will generally only last a few weeks, and may have longer or shorter durations depending on the challenge. The challenges will generally use fairly light-weight datasets that are synthetically generated from real-world data, and will provide an opportunity to quickly iterate through various model and feature engineering ideas, create visualizations, etc.

- **Synthetically-Generated Datasets**

Using synthetic data for Playground competitions allows us to strike a balance between having real-world data (with named features) and ensuring test labels are not publicly available. This allows us to host competitions with more interesting datasets than in the past. While there are still challenges with synthetic data generation, the state-of-the-art is much better now than when we started the Tabular Playground Series two years ago, and that goal is to produce datasets that have far fewer artifacts. Please feel free to give us feedback on the datasets for the different competitions so that we can continue to improve!

Case 3: House Prices - Advanced Regression Techniques

- *https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview*
- Goal: It is your job to predict the sales price for each house. For each Id in the test set, you must predict the value of the SalePrice variable.
  - Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.
  - With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.
- Metric: evaluated on Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price. (Taking logs means that errors in predicting expensive houses and cheap houses will affect the result equally.)

Case 4 : Location-based species presence prediction

- https://www.kaggle.com/competitions/geolifeclef-2025/overview
- Given GPS coordinates and various predictors, e.g., satellite images, climatic time series, land cover, human footprint, etc., a participant/team must predict a set of species that should grow there. To do so, we provide observation data comprising approximately 5 million Presence-Only (PO) occurrences and around 90 thousand Presence-Absence (PA) survey records
- **Motivation**
    - Predicting the plant species present at a given location is helpful for many biodiversity management and conservation scenarios. It allows for building high-resolution maps of species composition and related biodiversity indicators such as species diversity, endangered species, and invasive species. In scientific ecology, the problem is known as Species Distribution Modelling. Moreover, it could significantly improve the accuracy of species identification tools - such as Pl@ntNet - by reducing the list of candidate species observable at a given site.
    - More generally, it could facilitate biodiversity inventories by developing location-based recommendation services (e.g., on mobile phones), encouraging citizen scientist observers' involvement, and accelerating the annotation and validation of species observations to produce large, high-quality data sets.

Case 4 : Location-based species presence prediction

- https://www.kaggle.com/competitions/geolifeclef-2025/overview
- The evaluation metric for this competition is the samples-averaged F1-score, which measures an overlap between the predicted and actual set of species present at a given location and time.
- Each test PA sample i is associated with a set of ground-truth labels Yi, i.e., the set of plant species (=speciesId). For each sample, the submission must provide a list of labels, i.e., the set of species predicted present $Y\hat{}i$,1,$Y\hat{}i$,2,…,$Y\hat{}i$,Ri.
- The micro F1-score is then computed using

$$F_1 = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathrm{TP}_i}{\mathrm{TP}_i + (\mathrm{FP}_i + \mathrm{FN}_i)/2}$$

$$\text{Where} \begin{cases} \mathrm{TP}_i = \text{Number of predicted species truly present, i.e. } |\widehat{Y}_i \cap Y_i| \\ \mathrm{FP}_i = \text{Number of species predicted but absent, i.e. } |\widehat{Y}_i \setminus Y_i| \\ \mathrm{FN}_i = \text{Number of species not predicted but present, i.e. } |Y_i \setminus \widehat{Y}_i| \end{cases}$$

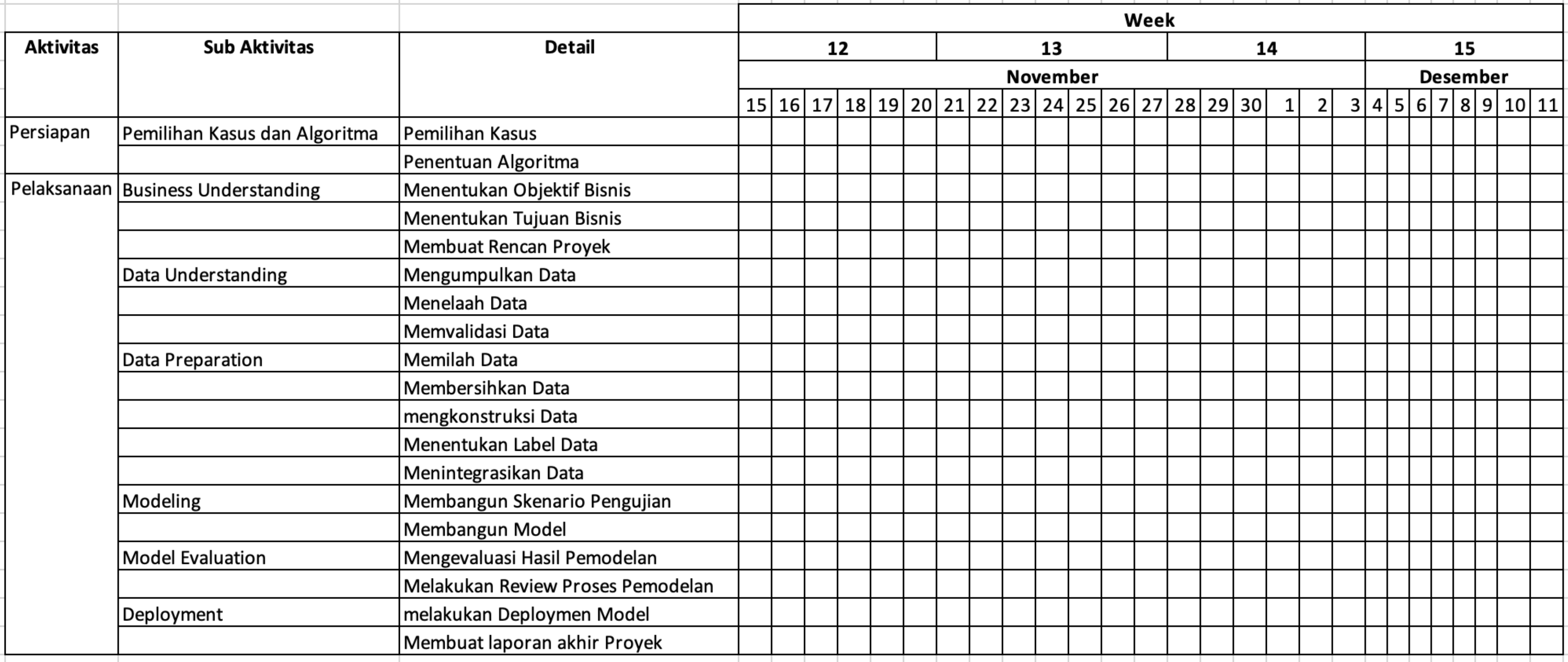
Case 5 : Stanford RNA 3D Folding (Solve RNA structure prediction, one of biology's remaining grand challenges)

- https://www.kaggle.com/competitions/stanford-rna-3d-folding/overview
- If you sat down to complete a puzzle without knowing what it should look like, you'd have to rely on patterns and logic to piece it together. In the same way, predicting Ribonucleic acid (RNA)'s 3D structure involves using only its sequence to figure out how it folds into the structures that define its function.
- In this competition, you'll develop machine learning models to predict an RNA molecule's 3D structure from its sequence. The goal is to improve our understanding of biological processes and drive new advancements in medicine and biotechnology.
- RNA is vital to life's most essential processes, but despite its significance, predicting its 3D structure is still difficult. Deep learning breakthroughs like AlphaFold have transformed protein structure prediction, but progress with RNA has been much slower due to limited data and evaluation methods.
- This competition builds on recent advances, like the deep learning foundation model RibonanzaNet, which emerged from a prior Kaggle competition. Now, you'll take on the next challenge—predicting RNA's full 3D structure.
- Your work could push RNA-based medicine forward, making treatments like cancer immunotherapies and CRISPR gene editing more accessible and effective. More fundamentally, your work may be the key step in illuminating the folds and functions of natural RNA molecules, which have been called the 'dark matter of biology'.
- This competition is made possible through a worldwide collaborative effort including the organizers, experimental RNA structural biologists, and predictors of the CASP16 and RNA-Puzzles competitions; Howard Hughes Medical Institute; the Institute of Protein Design; and Stanford University School of Medicine.

Progress

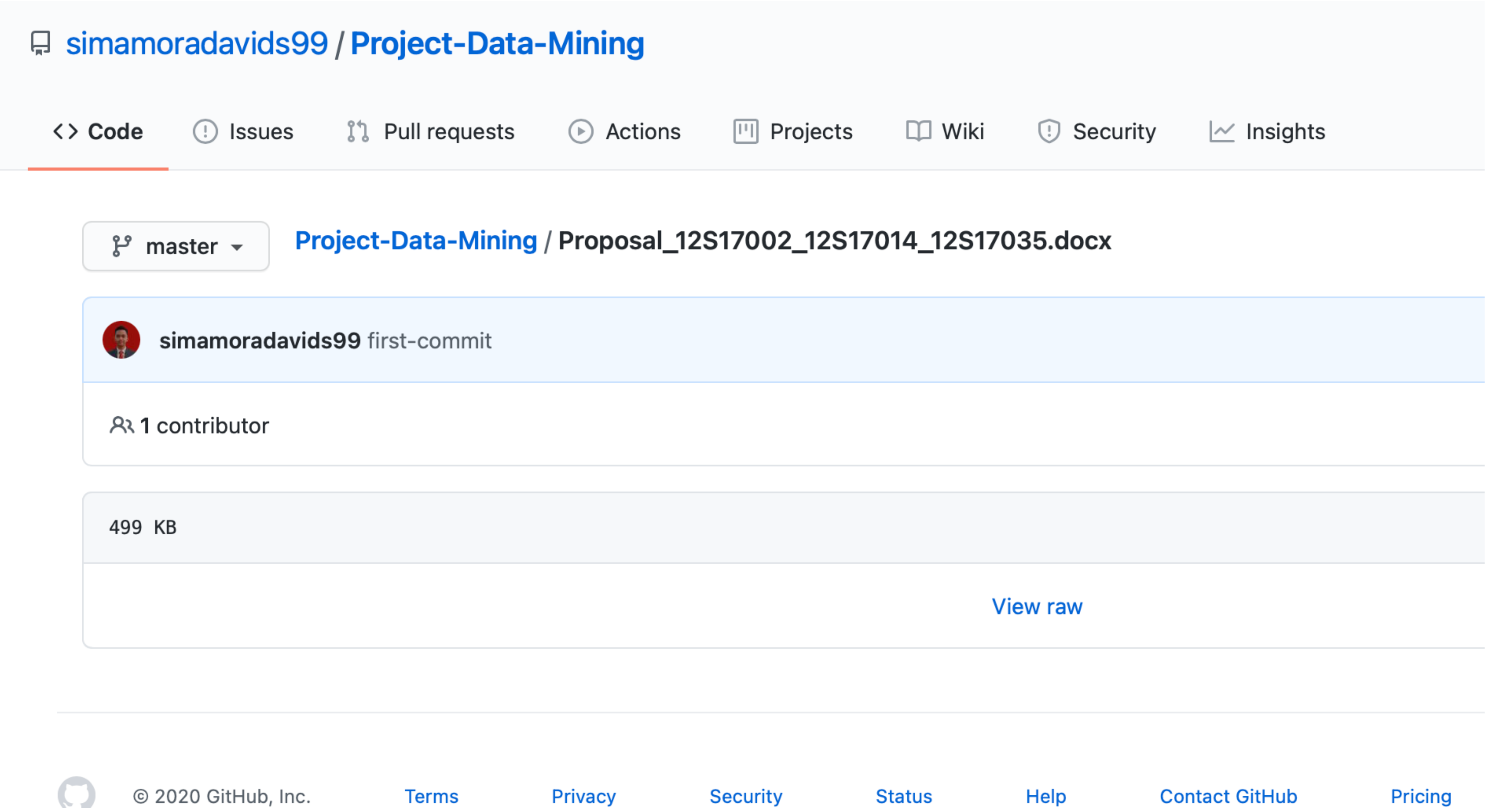
- Check Every Day or Week
  - Update TimeLine Every Day
  - Update Github Every Day
- Share Link for Google Sheet or Github

TimeLine



			Week																										
Aktivitas	Sub Aktivitas	Detail	12						13							14						15							
			November																			Desember							
			15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	1	2	3	4	5	6	7	8	9	10	11
Persiapan	Pemilihan Kasus dan Algoritma	Pemilihan Kasus																											
		Penentuan Algoritma																											
Pelaksanaan	Business Understanding	Menentukan Objektif Bisnis																											
		Menentukan Tujuan Bisnis																											
		Membuat Rencan Proyek																											
	Data Understanding	Mengumpulkan Data																											
		Menelaah Data																											
		Memvalidasi Data																											
	Data Preparation	Memilah Data																											
		Membersihkan Data																											
		mengkonstruksi Data																											
		Menentukan Label Data																											
		Menintegrasikan Data																											
	Modeling	Membangun Skenario Pengujian																											
		Membangun Model																											
	Model Evaluation	Mengevaluasi Hasil Pemodelan																											
		Melakukan Review Proses Pemodelan																											
	Deployment	melakukan Deploymen Model																											
		Membuat laporan akhir Proyek																											

Github



simamoradavids99 / Project-Data-Mining

Project-Data-Mining / Proposal_12S17002_12S17014_12S17035.docx

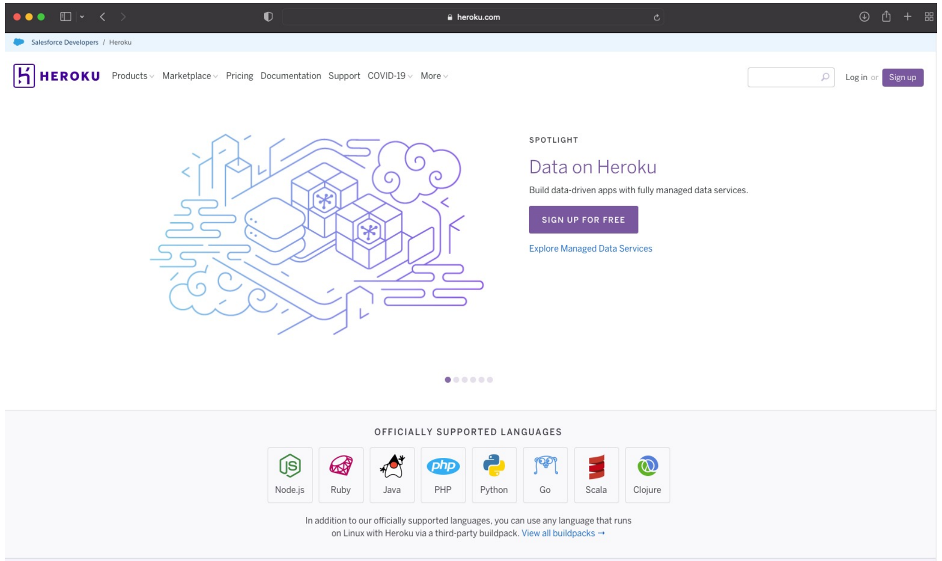
simamoradavids99 first-commit

1 contributor

499 KB

Deployment

- Deploy your Model on Cloud Application Platform
  - Heroku: platform as a service (PaaS) that enables developers to build, run, and operate applications entirely in the cloud. Provides free plan for students.



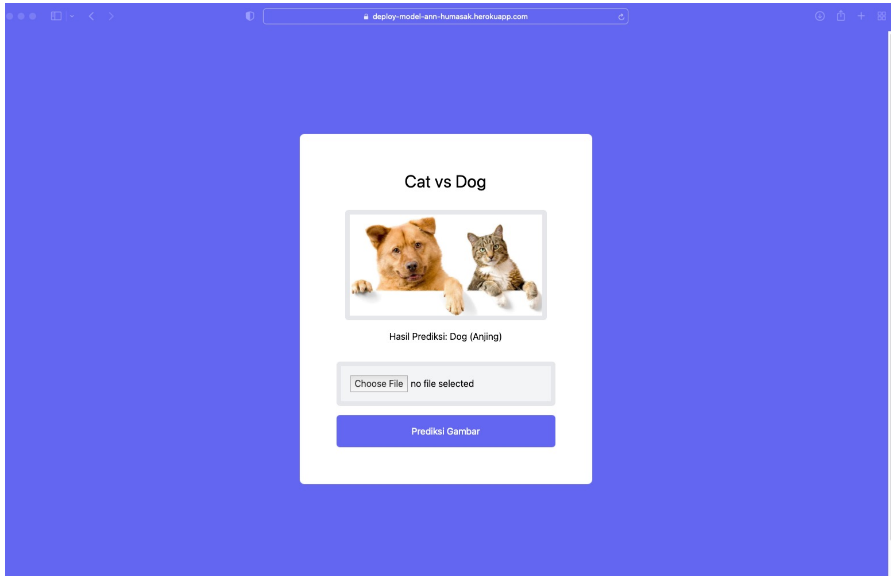
SPOTLIGHT

Data on Heroku

Build data-driven apps with fully managed data services.

OFFICIALLY SUPPORTED LANGUAGES

In addition to our officially supported languages, you can use any language that runs on Linux with Heroku via a third-party buildpack. View all buildpacks →



Cat vs Dog

Hasil Prediksi: Dog (Anjing)

https://deploy-model-ann-humasak.herokuapp.com

Constraints

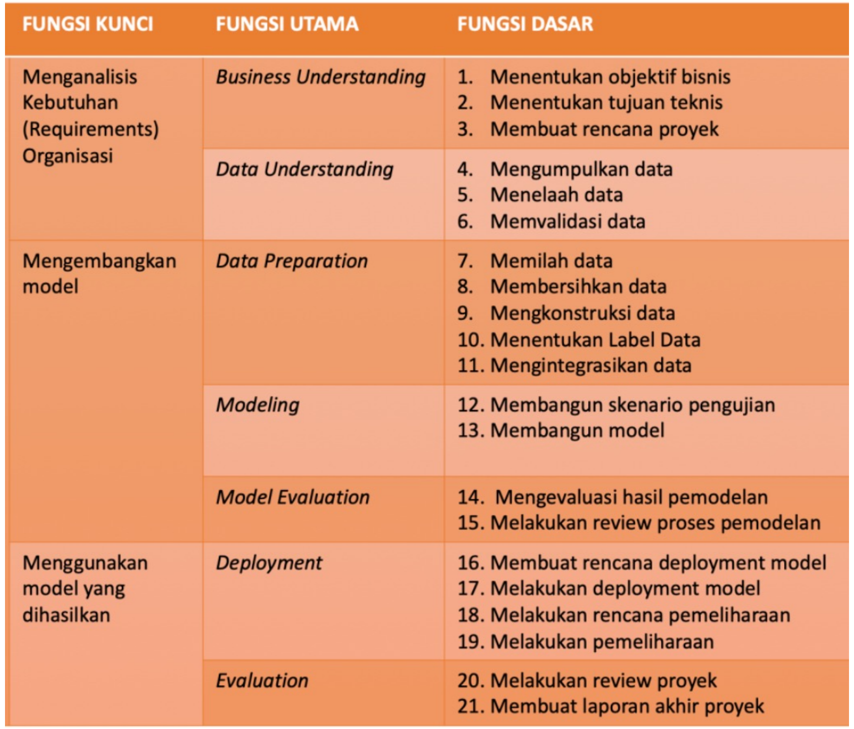
- Only two groups can choose same algorithm for each case
- Choose the best way for each step (or combine some techniques) to get higher (best) evaluation metrics
- You can skip one step only if you provide proof
- You can link data to the source in internet

Submission

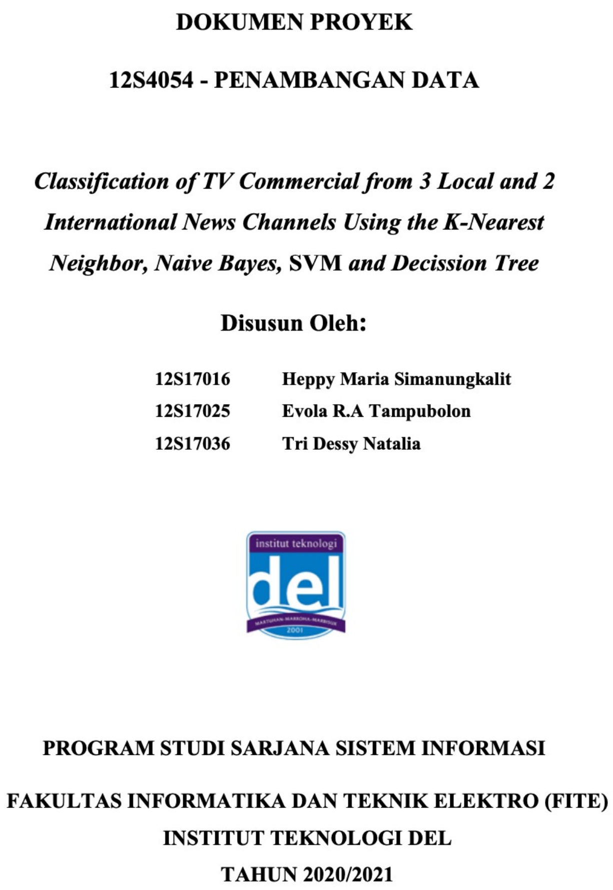
- Submit final report to https://ecourse.del.ac.id/.
- Create Timeline – per day : Green, Yellow, Red
- Github: per job done
- Final Report. Due date: Saturday, 3 May 2025 (22.00 WIB).
  - Document (Report)
  - Article draft
  - Application (Code and Deployment)
  - Poster
  - Video Project Presentation: Saturday, 3 May 2025 (22.00 WIB).
  - Deployment
- Deliverable Course: Case Book (Project) - 12S3205 Penambangan Data (Compiled by Ketua Kelas)

Submission

- Document/Report Structure :
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Model Evaluation
  - Deployment
- Final Report (Article draft)
  - Before submission, submit document to Turnitin, to check plagiarism.
  - The threshold of similarity < 15%



FUNGSI KUNCI	FUNGSI UTAMA	FUNGSI DASAR
Menganalisis Kebutuhan (Requirements) Organisasi	*Business Understanding*	1. Menentukan objektif bisnis 2. Menentukan tujuan teknis 3. Membuat rencana proyek
	Data Understanding	4. Mengumpulkan data 5. Menelaah data 6. Memvalidasi data
Mengembangkan model	*Data Preparation*	7. Memilah data 8. Membersihkan data 9. Mengkonstruksi data 10. Menentukan Label Data 11. Mengintegrasikan data
	Modeling	12. Membangun skenario pengujian 13. Membangun model
	Model Evaluation	14. Mengevaluasi hasil pemodelan 15. Melakukan review proses pemodelan
Menggunakan model yang dihasilkan	*Deployment*	16. Membuat rencana deployment model 17. Melakukan deployment model 18. Melakukan rencana pemeliharaan 19. Melakukan pemeliharaan
	Evaluation	20. Melakukan review proyek 21. Membuat laporan akhir proyek



DOKUMEN PROYEK

12S4054 - PENAMBANGAN DATA

Classification of TV Commercial from 3 Local and 2 International News Channels Using the K-Nearest Neighbor, Naive Bayes, SVM and Decission Tree

Disusun Oleh:

12S17016	Heppy Maria Simanungkalit
12S17025	Evola R.A Tampubolon
12S17036	Tri Dessy Natalia

PROGRAM STUDI SARJANA SISTEM INFORMASI

FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO (FITE)
INSTITUT TEKNOLOGI DEL
TAHUN 2020/2021

Grading

- $\text{Mark}_{\text{team}}$: 100
  - Progress : 20
  - Final Report : 20
  - Application : 60
- $\text{Mark}_{\text{individual}}$: 100
  - Presentation : 30
  - Contribution : 70
- $\text{Mark}_{\text{project}} = 0.5 * \text{Mark}_{\text{team}} + 0.5 * \text{Mark}_{\text{individual}}$

Advices

- Have the role of each member **very** well-defined form the beginning
- Agree on each single step before you start
- Use a task management app
- Choose a team leader
  - Has the right to have final decision when no agreement could be reached by members
  - Organizes work among members and follows progress

Advices

- If X can have outcome A
  team of $5X$ should have an outcome of $\gg 5A$
- Not Allowed:
  - Get a ready app/project and submit
  - Using collections that are not public
  - Plagiarism from other available sources (Kaggle, Github, etc) = 0
- Allowed:
  - Using libraries and tools
  - Discussing with other groups and sharing ideas

Today: 27 March 2025

- Choose case and algorithm (google sheet)
- Each group creates Github for Data Mining Project. Share and Open to HTS and Siska Manullang. Only group member, HTS, and Siska Manullang can access Github. Github structure consists of:
  - Document (Report)
  - Application (Code and Deployment)
  - Poster
  - Video Project Presentation
  - Deployment
- Every week progress will be checked through Github.
- Doing Business Understanding

Thank You

Colin Powell

"A dream does not become reality through magic; it takes sweat, determination, and hard work."