

# Final Project

12S3205 - Penambangan Data  
Semester Gasal 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 methodology  
(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	<i>Business Understanding</i>	1. Menentukan objektif bisnis 2. Menentukan tujuan teknis 3. Membuat rencana proyek
	<i>Data Understanding</i>	4. Mengumpulkan data 5. Menelaah data 6. Memvalidasi data
Mengembangkan model	<i>Data Preparation</i>	7. Memilah data 8. Membersihkan data 9. Mengkonstruksi data 10. Menentukan Label Data 11. Mengintegrasikan data
	<i>Modeling</i>	12. Membangun skenario pengujian 13. Membangun model
	<i>Model Evaluation</i>	14. Mengevaluasi hasil pemodelan 15. Melakukan review proses pemodelan
Menggunakan model yang dihasilkan	<i>Deployment</i>	16. Membuat rencana deployment model 17. Melakukan deployment model 18. Melakukan rencana pemeliharaan 19. Melakukan pemeliharaan
	<i>Evaluation</i>	20. Melakukan review proyek 21. Membuat laporan akhir proyek

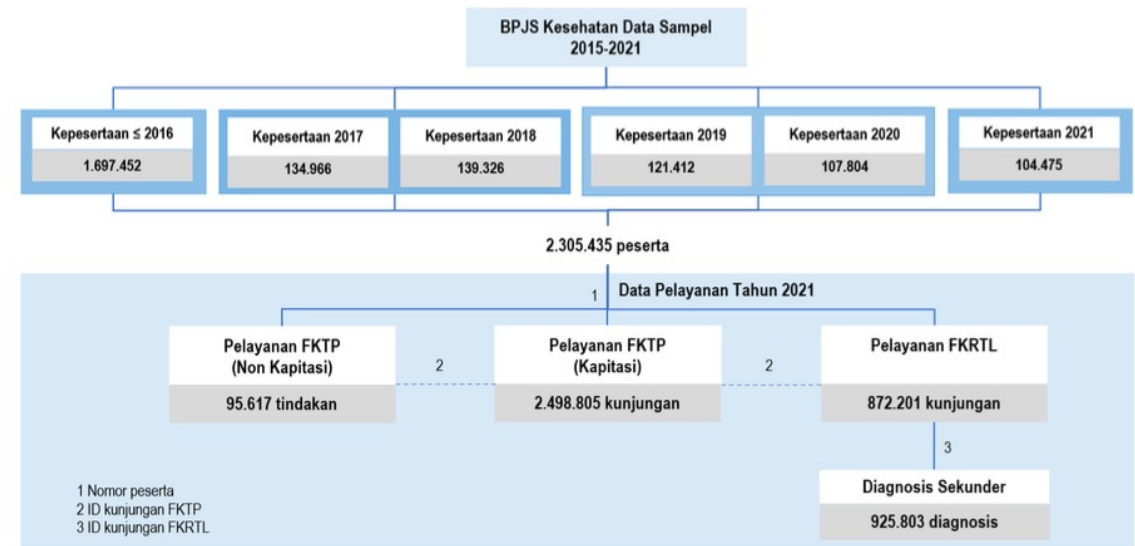
# Case: BPPJS Hackaton

- Case 1: Sample BPJS 2015 – 2021
- Case 2: Exploring Mental Health Data
- Case 3: House Prices - Advanced Regression Techniques
- Case 4: DiTenun: Ulos Image Classification
- Case 5: Kubis Image Detection

# 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/1eNDDeoaPfQJQWr5RUSVdgnobSXQ5VsKRU?usp=sharing>



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 : DiTenun - Ulos Image Classification

# Case 5 : Kubis Image Detection

# Progress

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

# TimeLine


Aktivitas	Sub Aktivitas	Detail	Week																																		
			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**

<> **Code**

! Issues

 Pull requests


▶ Actions

 Projects

 Wiki

 Security

 Insights

 master ▼

**Project-Data-Mining** / **Proposal\_12S17002\_12S17014\_12S17035.docx**



**simamoradavids99** first-commit

 1 contributor

499 KB

[View raw](#)



© 2020 GitHub, Inc.

[Terms](#)

[Privacy](#)

[Security](#)

[Status](#)

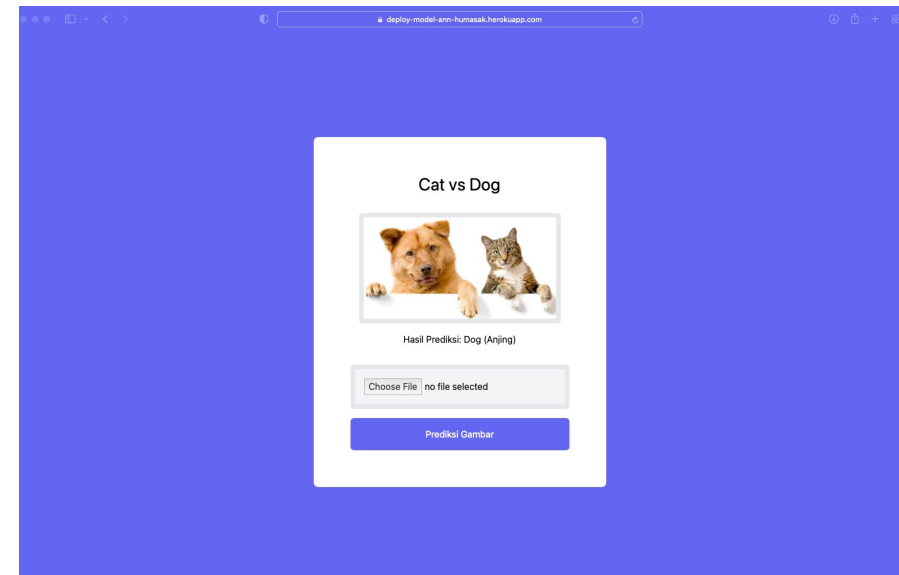
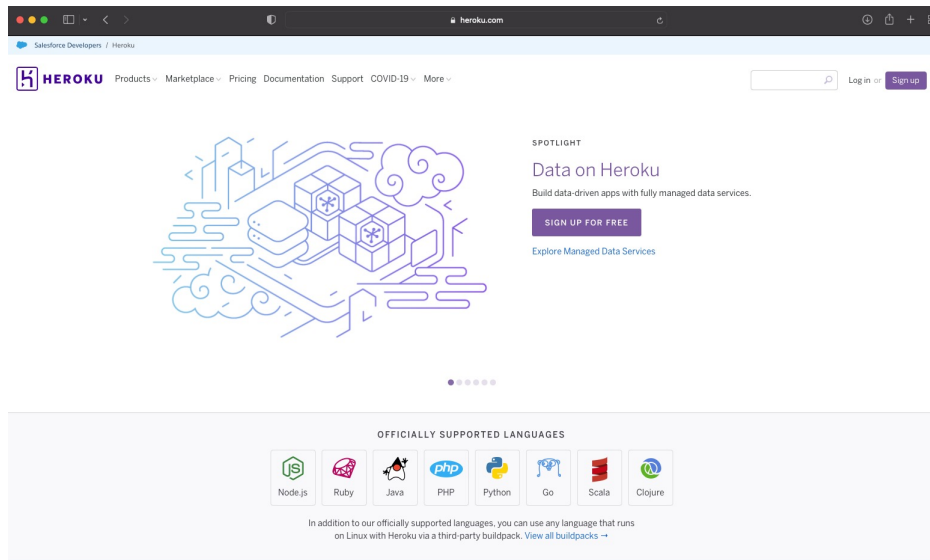
[Help](#)

[Contact GitHub](#)

[Pricing](#)

# 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.



<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, 14 December 2024** (22.00 WIB).
  - Document (Report)
  - Article draft
  - Application (Code and Deployment)
  - Poster
  - Video Project Presentation: **Saturday, 14 December 2024** (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	<i>Business Understanding</i>	1. Menentukan objektif bisnis 2. Menentukan tujuan teknis 3. Membuat rencana proyek
	<i>Data Understanding</i>	4. Mengumpulkan data 5. Menelaah data 6. Memvalidasi data
Mengembangkan model	<i>Data Preparation</i>	7. Memilah data 8. Membersihkan data 9. Mengkonstruksi data 10. Menentukan Label Data 11. Mengintegrasikan data
	<i>Modeling</i>	12. Membangun skenario pengujian 13. Membangun model
	<i>Model Evaluation</i>	14. Mengevaluasi hasil pemodelan 15. Melakukan review proses pemodelan
Menggunakan model yang dihasilkan	<i>Deployment</i>	16. Membuat rencana deployment model 17. Melakukan deployment model 18. Melakukan rencana pemeliharaan 19. Melakukan pemeliharaan
	<i>Evaluation</i>	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 Decision 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 from 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: 08 November 2024

- Choose case and algorithm (google sheet)
- Each group creates Github for Data Mining Project. Share and Open to HTS and Guntur Sinaga. Only group member, HTS, and Guntur Sinaga 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."