# TEORI - PENUGASAN OPEN RECRUITMENT
# DIVISI DATA SCIENCE DAN ARTIFICIAL INTELLIGENCE

## 2024/2025

| | |
|---|---|
| **Name** | **: Farhan Adiwidya Pradana** |
| **NIM** | **: 24/536804/PA/22773** |
| **Class** | **: CS A** |

1. Dengan menggunakan kalimat Anda sendiri, jelaskan yang dimaksud dengan EDA (Exploratory Data Analysis)! Mengapa hal tersebut penting dalam proses analisis data?

   Exploratory Data Analysis (Analisis Data Secara Keseluruhan) is an important step for data scientists when examining and investigating datasets in order to determine their key characteristics.

   EDA makes use of data visualization and statistical techniques to better comprehend patterns, find outliers, test hypotheses, and investigate correlations between variables.

   This technique assists data scientists in determining the most efficient way to prepare and manipulate data for meaningful analysis, ensuring that outcomes are consistent with business objectives.

2. Jelaskan perbedaan antara supervised learning, unsupervised learning, dan reinforcement learning! Termasuk kategori yang manakah problem set pada penugasan open recruitment ini?

   Differences between Supervised, Unsupervised, and Reinforcement Learning:

   A. Supervised Learning

      Supervised Learning is a Machine Learning approach that is defined by the use of labeled data sets. These data sets are designed to train or supervise algorithms to classify data or predict outcomes accurately. Supervised Learning uses a training dataset to develop a prediction model by consuming input data and output values. Examples of its applications include areas such as:

      - Image Recognition
      - Speech Recognition
      - Natural Language Processing

B. Unsupervised Learning

Unsupervised Learning is a type of Machine Learning that does not use labeled datasets (unlabeled data sets). These algorithms generally learn hidden structure patterns contained in the data where the desired output is unknown. The models learn to represent the underlying data structure without being explicitly told what to look for. Examples of its applications include areas such as:

- Computer Vision
- Medical Imaging
- Clustering

C. Reinforcement Learning

Reinforcement learning is a type of Machine Learning where a model learns through trial and error by interacting with the environment and then receiving a reward or penalty for its actions. Reinforcement Learning involves learning what to do-how to map a situation to an action-to maximize the numerical reward signal. Examples of its applications include areas such as:

- Robotics
- Game A.I (Chess engines, etc.)
- Autonomous Vehicles

The problem set in this open-recruitment assignment can be classified as Supervised Learning. But why?

- Labeled data

  Each row in the training dataset includes both input features (e.g., work_year, employment_type, remote_ratio, etc.) and the corresponding output or label (salary).
- Prediction task

The goal of the assignment is to predict salaries based on the input features in the test dataset. Which aligns with the typical/general objective of Supervised Learning that is predicting a target feature.

- Categories of supervised learning

  The problem given was a type of regression problem/regression task because the target variable (salary) is continuous.

3. Apa yang dimaksud dengan overfitting dan underfitting dalam konteks machine learning? Apakah dalam pengerjaan penugasan praktek Anda mengalami salah satu atau kedua masalah tersebut? Bagaimana Anda menanganinya?

   A. Overfitting

   Overfitting is an undesirable machine learning behavior that occurs when the model learns too well on the training data, capturing irrelevant details or noise. This results in the model having high performance on the training data but failing to generalize to new, unseen data.

   B. Underfitting

   Underfitting happens when a machine learning model does not learn the link between variables in the data, resulting in poor performance on both the training and validation/test datasets. This frequently occurs when the model is overly simplistic or lacks the necessary complexity to reflect the underlying patterns in the data.

   C. Scenario experienced

   During modeling, I implemented cross-validation before and after hyperparameter tuning using the Light Gradient Boosting Machine (LGBM) model. The local notebook results for RMSE, MAE, and R² were as follows:

   i. **Before Hyperparameter tuning (current time of writing)**

- **RMSE**: [56990.31, 56727.04, 58307.38, 57119.04, 56485.73]
- **Mean RMSE**: 57,125.90
- **Standard Deviation of RMSE**: 629.80
- **Mean MAE**: 44,654.62
- **Mean R²**: 0.3138
- **Testing RMSE**: 57,738.48

ii. **Final Lgbm performance (current time of writing)**

- **Testing RMSE**: 56,749.40
- **Testing MAE**: 44,490.33
- **Testing R²**: 0.3206

iii. **After Hyperparameter tuning (current time of writing)**

- **RMSE**: [56902.97, 56795.74, 58327.63, 57214.66, 56542.60]
- **Mean RMSE**: 57,156.72
- **Standard Deviation of RMSE**: 623.89
- **Mean MAE**: 44,694.05
- **Mean R²**: 0.3131

iv. **Public RMSE (best submission result-current)**

- **RMSE**: 60,121.02

The consistent cross-validation and testing RMSE scores suggest that the LGBM model performed well on the local test set. After hyperparameter tuning, the testing RMSE improved slightly, indicating that the tuning successfully refined the model.

However, the public RMSE score (evaluated on unseen data with a 30% split of the test set) was higher than the local RMSE, which suggests the following:

1. Possible overfitting:

   The model may have learned some patterns specific to the training and local test sets, leading to slightly reduced generalization on the public leaderboard data

2. Feature Distribution Differences:
   The unseen data in the public leaderboard might differ in distribution from the training and local test sets, making the predictions less accurate

**Steps taken to improve model performance (from previous submissions)**

a. Feature Engineering
   - Simplified skewed features like employment_type and company_size with ordinal encoding
   - Created a binary feature, new_work_own_country, indicating if the employee works in their country of residence.

b. Encoding methods
   - Target encoding for high-cardinality features (job_title, employee_residence, etc.)
   - Rare encoding to group infrequent categories.
   - One-hot encoding for remaining categorical features.

c. Regularization and Cross-Validation:
   - Used cross-validation to ensure the model could generalize well across different data splits
   - Applied hyperparameter tuning to strike a balance between underfitting and overfitting

While significant progress was made through those steps, there is still much more room for improvement.

4. Seandainya dalam proses prediksi penugasan problem set diperbolehkan menambahkan data eksternal, apakah Anda akan menggunakan data eksternal? Jika iya, data apa yang akan Anda gunakan dan jelaskan

alasannya! (NB: selain data primer harga laptop dengan spesifikasi yang sama, contoh: data harga laptop di marketplace)

After analyzing the data so far, there would be some more external data that I would add to assist in predicting salary:

- Industry trends: Salary benchmarks for the Data Science and Technology Industry, reports from Glassdoor or any other source could provide valuable context for salary ranges within specific job titles/roles. A data scientist in Great Britain might have a different salary range than one from in the Netherlands.
- Cost of living: Cost of living data for company_location and employee_residence can help contextualize salaries relative to living expenses in different regions. A salary of $60,000 might be considered high in one country but low in another with higher living costs.

5. Bagaimana tanggapan dan evaluasi Anda terhadap problem set pada penugasan praktek dan soal teori pada proses open recruitment ini?

   The problem set in this open recruitment process to me is really good for both practice and skill refinement. The design of it and the amount of unique/diverse features enabled me to learn more about visualizing features with high cardinality, and provided a realistic challenge often encountered in real-world data analysis tasks. The process of this open recruitment as a whole required careful consideration of feature engineering, encoding techniques, and model tuning.

# REFERENCES

https://www.ibm.com/topics/exploratory-data-analysis

Awad, M., & Khanna, R. (2015). *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Apress. https://link.springer.com/book/10.1007/978-1-4302-5990-9

Graves, A. (2013). *Generating Sequences With Recurrent Neural Networks*. arXiv. https://arxiv.org/abs/1308.0850

https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press. https://www.andrew.cmu.edu/course/10-703/textbook/BartoSutton.pdf

Campbell, M., Hoane Jr., A. J., & Hsu, F. (2002). Deep Blue. *Artificial Intelligence*. https://core.ac.uk/download/pdf/82416379.pdf

Bojarski, M., Tesla, D. D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., & Zieba, K. (2016). *End to End Learning for Self-Driving Cars*. arXiv. https://arxiv.org/abs/1604.07316

Kormushev, P., Calinon, S., & Caldwell, D. G. (2013). Reinforcement Learning in Robotics: Applications and Real-World Challenges†. *robotics*. 10.3390/robotics2030122OPEN

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. https://github.com/janishar/mit-deep-learning-book-pdf

algoritma. (n.d.). *2 Masalah Data Scientist: Overfitting & Underfitting*. algorit.ma. Retrieved November 16, 2024, from https://algorit.ma/blog/data-science/overfitting-underfitting/

Amazon AWS. (n.d.). *What is Overfitting? - Overfitting in Machine Learning Explained*. AWS. Retrieved November 16, 2024, from https://aws.amazon.com/what-is/overfitting/