

## **Project Plan – DATA ANALYTIC PROGRAMMING**

### **Active Member ID and Name :**

1. 2314020327 – **Jasmine Syarafina**
2. 2314020334 - **Rimta Shevanya Ginting**

### **Title of the Project :**

Airline Passenger Satisfaction Prediction Using Supervised Learning Algorithms

### **Executive Summary :**

Project ini akan menggunakan dataset penumpang maskapai penerbangan yang diperoleh dari platform Kaggle. Dataset tersebut terdiri dari dua file, yaitu "test.csv" dan "train.csv". Meskipun kedua file tersebut disediakan untuk memudahkan proses pemisahan data pelatihan dan pengujian, analisis ini hanya akan menggunakan file "train.csv" yang berisi 103.904 record dengan 25 atribut.

Persiapan Data Sebelum membangun model, beberapa langkah persiapan data perlu dilakukan. Pertama, variabel kategorik dan ordinal yang awalnya dikenali sebagai "integer" atau "chr" akan dikonversi menjadi tipe data faktor. Hal ini penting untuk memastikan variabel-variabel tersebut diperlakukan dengan benar selama pemodelan.

Selanjutnya, variabel target "Satisfaction" akan direkode ulang menjadi 0 (puas) dan 1 (netral/tidak puas). Pengkodean ulang ini dilakukan untuk menyederhanakan masalah klasifikasi menjadi biner, sehingga model dapat berfokus pada membedakan antara pelanggan puas dan tidak puas.

Penanganan data hilang juga akan menjadi bagian penting dalam persiapan data. Jika terdapat data hilang, akan diidentifikasi pola dan penyebabnya. Kemudian, akan dipertimbangkan metode imputasi yang sesuai untuk mengatasi data hilang tersebut.

Analisis Eksplorasi Data Setelah persiapan data selesai, akan dilakukan analisis eksplorasi data untuk memahami karakteristik dataset secara mendalam. Analisis yang akan dilakukan meliputi:

1. Proporsi variabel target "Satisfaction" untuk melihat apakah terdapat ketidakseimbangan kelas.
2. Analisis kepuasan berdasarkan tipe pelanggan (loyal vs tidak loyal), kelas perjalanan, dan tipe perjalanan (bisnis vs pribadi) untuk mengetahui pola-pola yang memengaruhi kepuasan.
3. Analisis korelasi antar variabel prediktor untuk mengidentifikasi hubungan dan potensi multikolinearitas.

Hasil analisis eksplorasi data ini akan digunakan untuk memahami dinamika data sebelum membangun model prediksi.

Implementasi Model Setelah memahami karakteristik dataset, langkah selanjutnya adalah membangun model prediksi kepuasan pelanggan. Dataset terlebih dahulu akan dibagi menjadi set pelatihan (80%) dan set pengujian (20%) dengan menggunakan teknik yang sesuai untuk memastikan hasil dapat direproduksi.

Setidaknya tiga jenis model akan diimplementasikan dan dievaluasi, yaitu Support Vector Machine (SVM), Random Forest, dan Artificial Neural Network (ANN). Masing-masing model akan dilatih dengan berbagai konfigurasi parameter dan kernel untuk mendapatkan performa terbaik.

Evaluasi model akan dilakukan dengan menggunakan metrik yang sesuai, seperti akurasi, Area Under the Curve (AUC), dan lain-lain. Hasil evaluasi akan dibandingkan untuk menentukan model terbaik yang akan dipilih untuk digunakan dalam memprediksi kepuasan pelanggan.

Selain itu, analisis faktor penting juga akan dilakukan untuk mengidentifikasi variabel-variabel yang paling berpengaruh terhadap kepuasan pelanggan. Informasi ini dapat berguna untuk memberikan wawasan bisnis dan strategi peningkatan kepuasan.

### **Dataset :**

Utilizing data from the Kaggle dataset at

<https://www.kaggle.com/datasets/teejmahal20/airline-passenger->

### **Paper/Journals/Articles :**

1. Jurnal penelitian tentang analisis kepuasan pelanggan maskapai penerbangan:
  - Bhaskar, H. L., & Ravi, V. (2011). A hybrid ANN-GA model for airline passenger satisfaction. *Production Planning & Control*, 22(7), 661-671.
  - Atalik, O., & Arslan, M. (2009). Customer Satisfaction of Airport Passenger Transportation Service Quality. *ICEIS 2009 - Proceedings of the 11th International Conference on Enterprise Information Systems*, Volume AIDSS, Milan, Italy, May 6-10, 2009, 291-298.
2. Jurnal tentang penggunaan machine learning untuk memprediksi kepuasan pelanggan:
  - Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2005). Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research*, 34(10), 2902-2917.
  - Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 218(1), 211-229.
3. Tutorial dan dokumentasi penggunaan algoritma machine learning populer:
  - Dokumentasi scikit-learn untuk Support Vector Machines (SVM): <https://scikit-learn.org/stable/modules/svm.html>
  - Dokumentasi scikit-learn untuk Random Forest: <https://scikit-learn.org/stable/modules/ensemble.html#random-forests>
  - Dokumentasi Keras/TensorFlow untuk Artificial Neural Networks (ANN): <https://www.tensorflow.org/guide/keras/overview>
4. Artikel dan sumber daya tentang persiapan data dan analisis eksplorasi:
  - "A Gentle Introduction to Exploratory Data Analysis" oleh Jason Brownlee: <https://machinelearningmastery.com/gentle-introduction-exploratory-data-analysis/>

- "Data Preprocessing in Machine Learning" oleh Towards Data Science: <https://towardsdatascience.com/data-preprocessing-in-machine-learning-d10e7834ea49>
- 5. Artikel dan sumber daya tentang pemilihan dan evaluasi model machine learning:
  - "How to Choose the Right Machine Learning Algorithm" oleh Analytics Vidhya: <https://www.analyticsvidhya.com/blog/2015/08/common-machine-learning-algorithms/>
  - "Evaluating Classification Models in Machine Learning" oleh Towards Data Science: <https://towardsdatascience.com/evaluating-classification-models-in-machine-learning-7a3c689a45a2>