# Airline Passenger Satisfaction Prediction Using Supervised Learning Algorithms

**Project Brief – DATA ANALYTIC PROGRAMMING**

**Active Member ID and Name :**

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

***Backgrounder*:**

Proyek ini bertujuan untuk memprediksi kepuasan penumpang maskapai dengan menggunakan algoritma *supervised learning*, khususnya **Support Vector Machine (SVM)**. Dataset yang digunakan berasal dari survei kepuasan penumpang maskapai yang mencakup 103.904 data dengan 23 atribut yang relevan. Proyek ini akan fokus pada eksplorasi, pembersihan data, serta implementasi model prediktif dengan evaluasi performa menggunakan metrik seperti akurasi, AUC, dan *confusion matrix*.

Tujuan utama proyek adalah:

1. **Mengidentifikasi faktor utama** yang memengaruhi kepuasan penumpang.
2. **Membangun model prediksi** menggunakan SVM dengan berbagai kernel (*Gaussian RBF*, *Linear*, *Polynomial*, dan *Sigmoid*).
3. **Membandingkan performa model SVM** berdasarkan kernel yang digunakan untuk menentukan kernel terbaik.

Dataset telah diproses untuk menghapus data yang redundan dan menangani nilai yang hilang, sehingga dapat menghasilkan analisis yang valid dan akurat. Hasil proyek ini diharapkan dapat memberikan wawasan tentang faktor-faktor penting yang memengaruhi kepuasan penumpang, serta model prediksi yang optimal.

***Planned execution:***

a. **Prapemrosesan Data**
   a. Menghapus atribut tidak relevan seperti X dan id.
   b. Mengonversi variabel ordinal dan kategorikal ke dalam format numerik (*dummy variables*).
   c. Menangani data yang hilang dengan metode imputasi (*imputation*).
   d. Normalisasi dan standarisasi variabel numerik untuk meningkatkan performa model.

b. **Eksplorasi Data**
   a. Visualisasi korelasi antar variabel untuk memahami hubungan penting.
   b. Analisis distribusi kepuasan berdasarkan tipe pelanggan, kelas perjalanan, dan tujuan perjalanan.

c. **Implementasi Model** Sebelum melatih model, dataset dibagi menjadi 2 bagian. Karena ukuran data yang sangat besar, dataset dibagi dalam rasio 80:20. 80% data masuk ke set pelatihan dan 20% data masuk ke set pengujian. set.seed() dari

paket "caTools" digunakan untuk memastikan dapat direproduksi hasil setiap kali kita melakukan proses pemisahan. Proses pemisahan telah berhasil diselesaikan.

a. **SVM**: Eksperimen dengan kernel berbeda (*Gaussian RBF*, *Linear*, *Polynomial*, dan *Sigmoid*).
b. Melakukan tuning parameter seperti cost C untuk meningkatkan performa model. Karena kernel Gaussian RBF memberikan hasil terbaik, maka kernel ini dipilih sebagai pilihan fungsi dalam fase penyetelan model. Selama fase penyetelan model, parameter biaya, C ditingkatkan dari 1 ke 3. Hal ini mengakibatkan peningkatan minimal dalam akurasi set pengujian dan pelatihan.

d. **Evaluasi Model**
   a. Menggunakan metrik seperti akurasi, AUC, dan *confusion matrix* untuk menilai performa model SVM.
   b. Membandingkan hasil antar kernel SVM untuk menentukan kernel terbaik berdasarkan hasil evaluasi.

***Hasil Evaluasi:***

Hasil evaluasi performa model SVM dengan berbagai kernel ditampilkan pada tabel berikut:

| **Type of Kernel** | **Accuracy Training Set** | **Accuracy Test Set** | **Model Fitness** | **Training Error** |
|---|---|---|---|---|
| Gaussian RBF (rbfdot) | 0.9622 | 0.9587 | Good Fit | 0.037811 |
| Linear (vanilladot) | 0.9361 | 0.9367 | Good Fit | 0.063893 |
| Hyperbolic Tangent Sigmoid (tanhdot) | 0.5685 | 0.5784 | Poor Fit | 0.431529 |
| Polynomial (polydot) | 0.9361 | 0.9367 | Good Fit | 0.063917 |
| Gaussian RBF with Cost Parameter = 3 (Tuned Model) | 0.9674 | 0.9614 | Good Fit | 0.0326 |

Dari hasil ini, kernel **Gaussian RBF dengan parameter Cost = 3** memberikan hasil terbaik dengan akurasi tertinggi pada data training (96.74%) dan test (96.14%), serta error training terkecil (0.0326). Kernel ini dipilih sebagai model terbaik untuk memprediksi kepuasan penumpang.

***Conclusion:***

Proyek ini berhasil memprediksi kepuasan penumpang maskapai dengan tingkat akurasi tinggi menggunakan algoritma Support Vector Machine (SVM). Analisis menunjukkan bahwa faktor utama yang memengaruhi kepuasan penumpang mencakup jenis pelanggan (loyalitas), kelas perjalanan, kenyamanan kursi, layanan dalam kabin, serta penanganan bagasi. Penggunaan SVM dengan kernel Gaussian RBF memberikan hasil terbaik, dengan akurasi mencapai 96.74% pada data pelatihan dan 96.14% pada data pengujian, serta error pelatihan terkecil sebesar 0.0326. Penyesuaian parameter Cost pada kernel ini berhasil meningkatkan performa model, menyeimbangkan

overfitting dan underfitting, sehingga menghasilkan generalisasi yang baik. Kernel lainnya, seperti Linear, Polynomial, dan Sigmoid, menunjukkan performa yang lebih rendah, dengan kernel Sigmoid menjadi yang terburuk. Hasil ini menegaskan efektivitas kernel Gaussian RBF dalam menangkap pola kompleks pada dataset. Proyek ini memberikan kontribusi nyata dalam pengelolaan pengalaman pelanggan maskapai dengan menyediakan model prediktif yang dapat digunakan untuk mengidentifikasi potensi masalah dan peluang perbaikan. Wawasan yang diperoleh dapat membantu maskapai meningkatkan layanan, menargetkan program loyalitas, dan meningkatkan efisiensi operasional. Untuk pengembangan lebih lanjut, disarankan memperluas dataset dengan data dari berbagai maskapai dan wilayah, menguji algoritma lain seperti Random Forest atau Neural Networks, serta mempertimbangkan analisis faktor temporal untuk mengevaluasi tren kepuasan pelanggan dari waktu ke waktu. Secara keseluruhan, proyek ini tidak hanya menghasilkan model prediktif yang andal tetapi juga memberikan wawasan strategis yang berpotensi mendukung pengambilan keputusan berbasis data di industri penerbangan.

---

## Lampiran

**Dataset Link**:
https://www.kaggle.com/datasets/teejmahal20/airline-passenger-

**Slide Presentation Link:**

**https://docs.google.com/presentation/d/11hhEmzqCzJJ_GpbLMerDesiQS3iJGvQz/edit?usp=share_link&ouid=112289244097076734664&rtpof=true&sd=true**

**Project Plan Link:**
https://drive.google.com/file/d/1C3ny44wPvqAvelNO_A8NQyFGQYwWRBpI/view?usp=sharing

---

## Referensi


1. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018
2. Hsu, C. W., Chang, C. C., & Lin, C. J. (2010). A practical guide to support vector classification. *Technical Report*, Department of Computer Science, National Taiwan University.
3. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. https://doi.org/10.1007/978-0-387-84858-7
4. Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

5. Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology, 24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565
6. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
7. Quinlan, J. R. (1996). Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research, 4*, 77–90. https://doi.org/10.1613/jair.279
8. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2nd ed.). O'Reilly Media.