Ekstraksi Informasi Teks dan Identifikasi Faktor Kecelakaan Pesawat dengan PySpark

*Textual Information Extraction and Identification of Factors in Airplane Accidents using PySpark*

**1Habibie Ed Dien\*, 2 Ony Novianti, 3Haarisah Yustika Putri Al-Jufri, 4Reynaldi Fakhri Pratama**

1,2,3,4Teknik Informatika, Jurusan Teknologi Informasi, Politeknik Negeri Malang

Jl. Soekarno Hatta No. 9 Kota Malang, Indonesia

\*e-mail: [*habibie@polinema.ac.id*](mailto:habibie@polinema.ac.id)

(***received*:** ?, ***revised*:** ?, ***accepted*:** ? diisi oleh editor)

**Abstrak**

Paper ini membahas penggunaan model Latent Dirichlet Allocation (LDA) dan algoritma klasifikasi Support Vector Classifier (SVC) untuk mengelompokkan faktor-faktor yang berkontribusi terhadap kecelakaan pesawat. Dataset yang terdiri dari informasi kecelakaan pesawat, termasuk kolom Summary, digunakan dalam penelitian ini. Langkah pertama adalah pra-pemrosesan data untuk membersihkan dan mempersiapkan dataset. Selanjutnya, model LDA diterapkan untuk mengidentifikasi topik-topik yang ada dalam kolom Summary. Distribusi topik yang dihasilkan oleh model LDA digunakan sebagai fitur-fitur dalam klasifikasi. Model SVC dilatih menggunakan fitur-fitur tersebut dan label kelas dari dataset latih. Performa model dievaluasi menggunakan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score. Penelitian ini juga mengeksplorasi penyetelan parameter pada model LDA dan SVC untuk meningkatkan performa. Hasil penelitian ini dapat digunakan untuk mengidentifikasi faktor-faktor yang berkontribusi terhadap kecelakaan pesawat dan meningkatkan keamanan penerbangan melalui tindakan pencegahan yang tepat.

**Kata kunci:** Pengelompokan, Faktor Kecelakaan Pesawat, LDA, SVC, Kecelakaan Pesawat

***Abstract***

*This paper discusses the utilization of the Latent Dirichlet Allocation (LDA) model and the Support Vector Classifier (SVC) algorithm to cluster the contributing factors to airplane accidents. A dataset consisting of airplane accident information, including the Summary column, is used in this research. The first step involves data preprocessing to clean and prepare the dataset. Next, the LDA model is applied to identify the topics present in the Summary column. The topic distributions generated by the LDA model are used as features in the classification process. The SVC model is trained using these features and the class labels from the training dataset. The performance of the model is evaluated using metrics such as accuracy, precision, recall, and F1-score. This study also explores parameter tuning for the LDA and SVC models to enhance their performance. The findings of this research can be utilized to identify the factors contributing to airplane accidents and improve aviation safety through appropriate preventive measures.*

***Keywords:*** *Clustering, Aircraft Accident Factors, LDA, SVC, Aircraft Accidents*

# Pendahuluan (or Introduction)

Sektor transportasi merupakan salah satu sektor yang berperan penting dalam perekonomian suatu wilayah, baik dalam skala kecil maupun besar. Karena bersifat menghubungkan suatu daerah dengan daerah lain, kecelakaan transportasi melalui jalur darat, laut, maupun udara akan selalu mendapat perhatian yang besar dari masyarakat luas. Terutama kecelakaan pada sektor transportasi udara. [[1]](https://www.zotero.org/google-docs/?89rqHu) Pesawat merupakan moda transportasi yang pengguna jasanya merupakan masyarakat berpenghasilan menengah keatas, sehingga terkesan eksklusif dan berteknologi tinggi. Karena itu kecelakaan pesawat akan lebih menarik perhatian publik.

Ada dua istilah yang mewakili peristiwa tidak sempurnanya penerbangan, yaitu kejadian (incident) dan kecelakaan (accident). Kejadian mengacu pada kejadian selama penerbangan yang berkaitan dengan pengoperasian pesawat dan berpotensi membahayakan keselamatan selama penerbangan. Sedangkan kecelakaan mengacu pada peristiwa luar biasa di luar kemampuan manusia yang terjadi selama penerbangan dan berdampak pada kerugian, luka parah, hingga kematian yang menimbulkan kerusakan struktural. [[2]](https://www.zotero.org/google-docs/?9Aqr9t) Keduanya sama-sama menimbulkan kerugian sehingga tentu saja kecelakaan tersebut sebaiknya dihindari.

Oleh karena itu, pada penelitian ini kami bertujuan untuk mengidentifikasi faktor-faktor penyebab terjadinya kecelakaan pesawat dengan harapan penyebab tersebut dapat dihindari sebaik mungkin agar tidak menimbulkan kerugian.

# Tinjauan Literatur (or Literature Review)

Damien Kelly dan Marina Efthymiou melaksanakan sebuah studi yang berjudul “*An analysis of human factors in fifty controlled flight into terrain aviation accidents from 2007 to 2017*”[[1]](https://www.zotero.org/google-docs/?cT70aK). Tujuan dari penelitian ini adalah untuk mengidentifikasi faktor-faktor yang berperan dalam 50 kasus kecelakaan CFIT (Controlled Flight Into Terrain) di 24 negara selama periode 10 tahun tersebut. Metode yang digunakan dalam penelitian ini adalah HFACS (Human Factors Analysis and Classification).

Sedangkan pada studi yang dilakukan oleh Adya Lucyana, Rossi Passarella, Dedy Kurniawan, Purwita Sari, dan Muhammad Ali Buchari yang berjudul “Analisis Penyebab Kecelakaan Pesawat di Indonesia Menggunakan Metode K-Means”[[2]](https://www.zotero.org/google-docs/?k7rd0w), ditemukan bahwa 44% dari kecelakaan CFIT terjadi selama penerbangan pelayaran (cruise flight), di mana faktor-faktor manusia memainkan peran utama dalam fase ini. Hal ini disebabkan oleh perlunya menjaga stabilitas pesawat selama penerbangan yang aman. Gangguan, rasa angkuh terhadap kemampuan diri sendiri, dan kelelahan adalah beberapa faktor yang mungkin dirasakan oleh kru penerbangan sebagai kontributor terhadap kecelakaan CFIT selama fase penerbangan tersebut. Salah satu penyebab lain dari kecelakaan CFIT adalah ketiadaan perangkat peringatan terhadap medan seperti EGPWS dan GPWS, terutama pada pesawat terbang general aviation yang umumnya tidak dilengkapi dengan perangkat tersebut. Hal ini dikarenakan biaya tinggi yang terkait dengan pembelian dan pemasangan perangkat tersebut, terutama karena sebagian besar pesawat jenis ini digunakan secara pribadi. Selain itu, tidak ada persyaratan internasional yang mengharuskan pemasangan perangkat ini pada jenis pesawat tersebut, berbeda dengan penerbangan komersial yang mewajibkan penggunaan perangkat tersebut pada pesawat mereka.

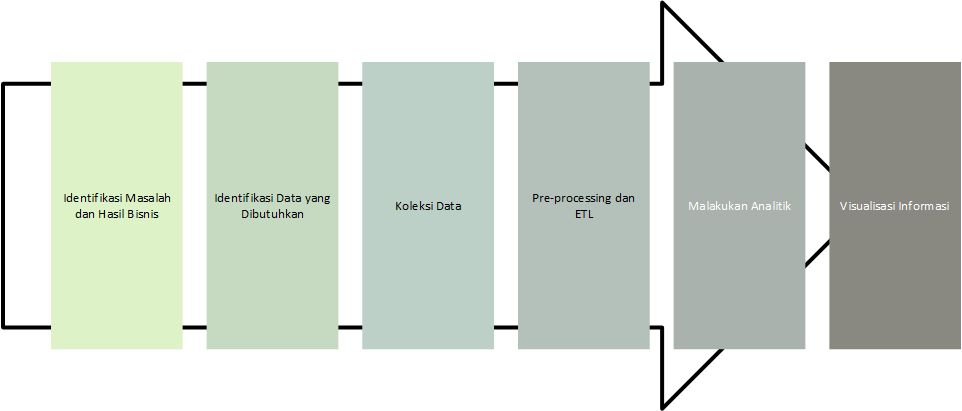
Dalam penelitian ini, digunakan metode K-Means untuk mengelompokkan data, dan untuk menentukan jumlah cluster yang optimal, digunakan metode elbow yang menghasilkan jumlah cluster K yang sama pada berbagai jumlah data yang berbeda. Hasil dari uji cluster menunjukkan bahwa jumlah cluster terbaik adalah 2 cluster dengan nilai elbow K = 2 sebesar 8000. Setiap cluster memiliki jumlah data yang berbeda, dengan cluster 1 memiliki 465 data dan cluster 2 memiliki 60 data.

Melalui analisis pada cluster 1, ditemukan bahwa deskripsi data untuk variabel PC termasuk dalam kategori homogen dengan nilai dispersi sebesar 0,67, sedangkan untuk variabel Total Fatalities dan Total Injured termasuk dalam kategori heterogen dengan nilai dispersi masing-masing sebesar 2,083 dan 3,924. Selanjutnya, untuk variabel PAX, termasuk dalam kategori homogen dengan nilai dispersi sebesar 0,27.

Sementara itu, analisis pada cluster 2 menunjukkan bahwa deskripsi data untuk variabel PC termasuk dalam kategori homogen dengan nilai dispersi sebesar 0,61, sedangkan untuk variabel Total Fatalities dan Total Injured termasuk dalam kategori heterogen dengan nilai dispersi masing-masing sebesar 3,611 dan 4,251. Untuk variabel PAX, termasuk dalam kategori homogen dengan nilai dispersi sebesar 0,95.

Dalam distribusi data antara cluster 1 dan cluster 2, tingkat keparahan kecelakaan pesawat didominasi oleh kategori ringan, dengan faktor penyebab terbanyak adalah Human. Selanjutnya, diikuti oleh kategori parah, dengan faktor penyebab terbanyak adalah Machines.

# Metode Penelitian (or Research Method)

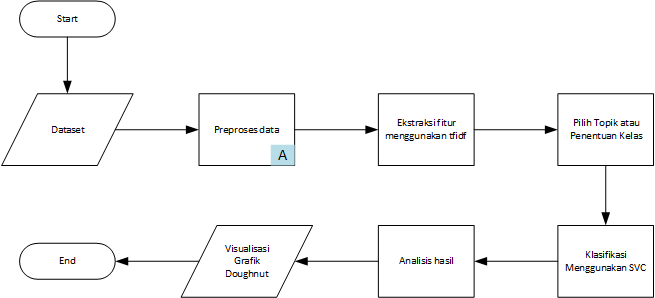


Studi mengenai penyebab kecelakaan pesawat memiliki pentingnya yang besar karena implikasinya yang luas terhadap keselamatan penerbangan dan kesejahteraan manusia.[[3]](https://www.zotero.org/google-docs/?hVMiGx) Setiap kecelakaan memberikan pelajaran berharga yang membantu mengidentifikasi kelemahan potensial dalam sistem pesawat, prosedur operasional, dan faktor manusia. Investigasi yang teliti memungkinkan para ahli untuk menentukan alasan mendasar di balik kecelakaan, seperti kegagalan mekanis, kekurangan desain, kesalahan pilot, atau pengaruh eksternal. Dengan memeriksa data kecelakaan secara cermat, otoritas penerbangan, insinyur, dan regulator dapat mengembangkan strategi yang efektif untuk meningkatkan standar keselamatan, memajukan teknologi, dan menerapkan program pelatihan komprehensif. Penelitian ini memungkinkan pengambilan tindakan proaktif, mencegah kecelakaan di masa depan dan meminimalkan kehilangan nyawa yang tidak bersalah. Selain itu, studi kecelakaan pesawat mendorong budaya pembelajaran yang berkelanjutan, di mana temuan-temuan berkontribusi pada pengembangan praktik dan standar keselamatan di seluruh industri penerbangan. Pada akhirnya, pemahaman terhadap penyebab kecelakaan pesawat sangat penting untuk kemajuan yang berkelanjutan dan keselamatan mutlak perjalanan udara.

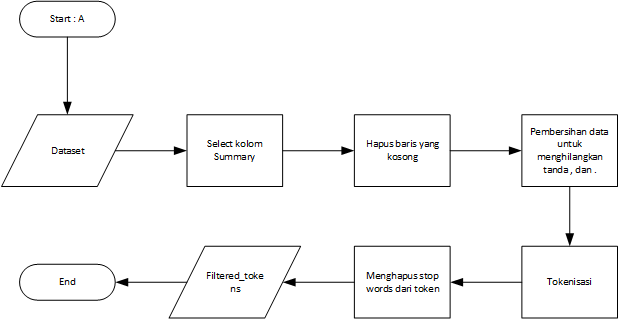
Untuk mendapatkan pemahaman tersebut, dibutuhkan data-data yang mendukung yaitu data-data mengenai informasi kecelakaan pesawat yang terjadi di seluruh dunia. Minimal data-data tersebut mencakup informasi mengenai jenis pesawat, jumlah penumpang, dan penyebab kecelakaan pesawat. Dalam penelitian ini, kami menggunakan metode analisis teks untuk mencari faktor-faktor apa saja yang paling banyak menyebabkan kecelakaan pesawat. Data yang akan kami olah, kami peroleh dari https://www.kaggle.com/datasets/saurograndi/airplane-crashes-since-1908. Kami memanfaatkan kolom summary untuk mengidentifikasi faktor penyebab kecelakaan pesawat dengan menghitung kata terbanyak atau kata yang sering muncul. Sebagai hasil, kami mendapatkan jumlah kata-kata yang muncul dalam setiap barisnya dan melakukan analisis sehingga mengetahui kata apa yang berpotensi menjadi faktor penyebab kecelakaan pesawat.

Dataset yang telah kami sebutkan sebelumnya merupakan kumpulan data mengenai kecelakaan pesawat dari tahun 1908 hingga 2009. Data ini berisi beberapa informasi penting seperti tanggal, waktu, lokasi, operator, jenis pesawat, rute penerbangan, jumlah penumpang, jumlah kematian, dan deskripsi singkat tentang kecelakaan. Diharapkan dengan mengolah data tersebut, kita dapat mengetahui faktor-faktor penting apa yang menyebabkan kecelakaan pesawat. Untuk hasil yang lebih maksimal, hasil tersebut diharapkan dapat diolah lagi sehingga dapat menghasilkan informasi mengenai area-area tertentu yang memerlukan perhatian lebih dalam upaya meningkatkan keselamatan penerbangan.

Kami memanfaatkan Apache Spark dan modul PySpark untuk mengolah data. Dengan menggunakan teknologi big data seperti Apache Spark, kami dapat mengelola data dengan volume yang besar dengan cepat untuk mendapatkan informasi penting dalam rangka meningkatkan keamanan dan keselamatan pada bidang penerbangan. Beberapa tahapan yang kami lakukan untuk mendapatkan informasi mengenai faktor kecelakaan pesawat dapat dilihat pada flowchart dibawah ini.



*Gambar 1 Flowchart*

**

*Gambar 2 Flowchart preprocess data*

Pada tahap awal kami mengambil data kecelakaan pesawat dari Kaggle, data tersebut kami unduh dalam bentuk atau format csv. Selanjutnya kami meload dataset tersebut ke dalam Spark sehingga dapat kami olah disana. Setelah data berhasil di load, kami memilih kolom Summary yang akan kami hitung jumlah kemunculan kata tertentu untuk menentukan faktor-faktor apa saja yang menyebabkan kecelakaan pesawat. Langkah selanjutnya adalah melakukan penghapusan baris kosong dari kolom summary agar tidak mengganggu hasil analisis. Setelah baris kosong berhasil dihapus, data dibersihkan dengan cara menghilangkan karakter-karakter tanda baca seperti “.”, “,”. Kemudian dilakukan tokenisasi pada data yang telah dibersihkan, dan langkah terakhir menghapus stopwords dari token.

Setelah berhasil melakukan *preprocessing*, token-token yang didapat digabungkan menjadi 1 dengan tujuan memberikan bobot pada kata-kata yang telah didapatkan menggunakan *library* fit\_transform. Data yang telah melalui pre-processing kami bagi menjadi dua bagian dengan perbandingan 80% untuk *data training* dan 20% untuk *data testing*. Kemudian dilakukan pemodelan menggunakan SVC untuk mengklasifikasikan masing-masing kata, lalu dilakukan prediksi data. Setelah semua selesai, kami menampilkan hasilnya dalam bentuk *pie chart* untuk menunjukkan proporsi jumlah masing-masing penyebab kecelakaan pesawat.

# Hasil dan Pembahasan (or Results and Analysis)

Hasil penelitian ini menunjukkan bahwa penggunaan model LDA dalam mengidentifikasi topik-topik terkait faktor kecelakaan pesawat berdasarkan kolom Summary memberikan wawasan yang berharga. Melalui analisis distribusi topik, kami dapat mengidentifikasi kelompok faktor yang sering muncul menjadi 9 kelas, yaitu :

1. Kelas Kecelakaan Pesawat di Rute Penerbangan
2. Kelas Kecelakaan Pesawat Akibat Mengambil atau Mendarat di Daerah Pegunungan
3. Kelas Kecelakaan Pesawat saat Pendekatan Landasan atau Mendarat
4. Kelas Kecelakaan Helikopter akibat Serangan Pemberontak
5. Kelas Kecelakaan Pesawat saat dalam Penerbangan (Masalah Internal)\
6. Kelas Kecelakaan Pesawat Segera Setelah Lepas Landas
7. Kelas Kecelakaan Pesawat saat Mendarat atau Lepas Landas
8. Kelas Kecelakaan Pesawat saat Lepas Landas di Daerah Berbukit atau Akibat Cuaca
9. Kelas Kecelakaan Pesawat saat Penerbangan

Pemanfaatan model SVC dalam klasifikasi faktor-faktor kecelakaan pesawat berdasarkan fitur-fitur yang dihasilkan dari model LDA menghasilkan hasil yang menjanjikan. Model SVC mampu mengklasifikasikan faktor-faktor kecelakaan pesawat dengan tingkat akurasi yang tinggi, yang memberikan keandalan dalam mengidentifikasi penyebab kecelakaan.

Penelitian ini juga membahas penyetelan parameter pada model LDA dan SVC. Hasil eksperimen menunjukkan bahwa penyetelan parameter yang optimal pada kedua model dapat meningkatkan performa klasifikasi. Sebagai contoh, penyesuaian jumlah topik pada model LDA dapat menghasilkan hasil yang lebih baik dalam mengidentifikasi topik-topik yang relevan dengan faktor kecelakaan pesawat. Selain itu, penyetelan parameter seperti parameter kernel pada SVC juga dapat mempengaruhi performa klasifikasi, dengan pemilihan kernel yang sesuai dapat meningkatkan pemisahan antar kelas.

Hasil penelitian ini memberikan kontribusi penting dalam pemahaman faktor-faktor kecelakaan pesawat dan memberikan landasan untuk meningkatkan keamanan penerbangan. Dengan memanfaatkan hasil pengelompokan dan klasifikasi ini, langkah-langkah pencegahan yang tepat dapat diambil untuk mengurangi risiko kecelakaan pesawat di masa depan. Selain itu, temuan ini dapat digunakan oleh perusahaan penerbangan, otoritas penerbangan, dan badan regulasi untuk mengarahkan upaya mereka dalam meningkatkan keselamatan penerbangan melalui langkah-langkah yang lebih terfokus dan efektif.

# Kesimpulan (or Conclusion)

Dalam penelitian ini, kami berhasil menerapkan model Latent Dirichlet Allocation (LDA) untuk mengidentifikasi topik-topik yang terkait dengan faktor-faktor kecelakaan pesawat berdasarkan kolom Summary. Distribusi topik yang dihasilkan oleh model LDA kemudian digunakan sebagai fitur-fitur dalam proses klasifikasi menggunakan Support Vector Classifier (SVC). Melalui evaluasi performa menggunakan metrik evaluasi yang relevan, kami mengamati tingkat akurasi, presisi, recall, dan F1-score yang memadai dari model yang dilatih. Penelitian ini menghasilkan pemahaman yang lebih baik tentang faktor-faktor yang berkontribusi terhadap kecelakaan pesawat dan memberikan dasar untuk meningkatkan keamanan penerbangan melalui tindakan pencegahan yang tepat. Dengan memanfaatkan hasil pengelompokan dan klasifikasi ini, perusahaan penerbangan dan otoritas terkait dapat mengidentifikasi area-area yang memerlukan perhatian lebih dan mengambil langkah-langkah yang diperlukan untuk mengurangi risiko kecelakaan pesawat di masa depan.

# Referensi (Reference)

An analysis of human factors in fifty controlled flight into terrain aviation accidents from 2007 to 2017. (2019). Journal of Safety Research, 69, 155–165. https://doi.org/10.1016/j.jsr.2019.03.009

ANALISIS PENYEBAB KECELAKAAN PESAWAT DI INDONESIA MENGGUNAKAN METODE K-MEANS. (2022). Jusikom : Jurnal Sistem Komputer Musirawas, 7(2), 89–105. https://doi.org/10.32767/jusikom.v7i2.1730

Pakan, W. (2008). Faktor Penyebab Kecelakaan Penerbangan di Indonesia Tahun 2000-2006. WARTA ARDHIA, 34(1), Article 1.

Puspitasari, R., & Giyono, U. (2021). KAJIAN KESETARAAN GENDER DAN PERLINDUNGAN ANAK TERHADAP KORBAN KECELAKAAN LALU LINTAS UDARA (Studi kasus pada Perempuan dan Anak pada Kecelakaan Pesawat Terbang Sriwijaya Air SJ 182). Edueksos Jurnal Pendidikan Sosial & Ekonomi, 10(1), Article 1. https://www.syekhnurjati.ac.id/jurnal/index.php/edueksos/article/view/7987