**TDS – Final Research Project**

Gal Pearl, Chen Shein

**I – Abstract**

Flight delays are a significant issue in air transport, which is impacting passengers, airlines, operations, causing financial losses as well. Accurate prediction of delays requires analyzing different factors that contribute to delays, such as departure times, the performance of companies & airports, weather conditions and more.   
In this project we applied Frequent Pattern Mining techniques, such as Association Rule Mining (Apriori Rules) & FP – Growth, to identify hidden relationships between those attributes and delays. The patterns are then used to generate new predictive features, which we incorporated into ML models. To evaluate the new features’ impact, we used XGBoost and Random Forest models, on four different flight datasets. In the early stages, we achieved weak correlations (0.19), which gave us an indication of the complexity of the delay factors. When we switched to non-linear models (Random Forest & XGB), the results were much better, and we noticed an improvement in prediction accuracy. Analyzing Feature Importance revealed that departure time, airline, and airport location are the most influential factors in delay prediction.  
Despite our attempted improvements, our model still faces the challenge due to stochastic natures of flight delays. We can enhance the model further using real-time weather condition tracking & air traffic/airport congestion, which are out of the scope of the course.

**II - Problem Description**

We focused on improving Pattern Mining in the DS pipeline, in context of flight delay prediction. We used the techniques beforementioned to discover hidden relationships between different attributes.

Improvement is required in order to defeat the struggle with capturing complex dependencies between different, multiple factors. Approaches such as linear regression or statistical analysis usually fail to reveal non-linear patterns and small correlations between attributes. It’s limiting the ability to generate insights and construct strong predictive models.

Therefore, we implemented several changes in our approach. First, we showed low correlations in linear models, indicating that these models aren’t sufficient for modelling flight delays; Second, we used high dimensional & complex interactions, as delays are influenced by multiple features and factors that we described above and more. Identifying those frequent patterns in this complex environment isn’t an easy task; Third, there was a lack in traditional models to incorporate pattern-based feature engineering.

We aim to generate better predictive features using pattern mining, to lead to a more accurate flight delays model.

**III - Our Solution**

We integrated FPM with ML models in order to improve flight delay predictions, in two stages:  
1) Apriori Algorithm to find association rules and FP – Grown to extract meaningful relationships between the attributes and the delays. First we made sure that the fields are matching in name and data, meaning we discarded from any early arrivals and delays which are less than 15 minutes, transforming continuous and categorical attributes into a format that suits pattern mining. Next, we extracted frequent itemsets that indicated strong associations with delays. Lastly, we generated new features, our discovered patterns that converted into predictive features, such as – high delay likelihood, a flag to indicate airline-time-airport combination, and pattern based categorical encoding that reflects flight delay trends.

2) Integrating with ML model: We tested the new features, and how they improve the prediction performance by incorporating them into XGB and RF models. We trained the baseline models in XGB and RF on the original datasets for a baseline performance. Next, we trained with the pattern based features we mentioned above, and compare the performance between the models, and evaluate the improvements using known metrics (Mean Absolute Error, RMSE and ). We found that linear models generate weak correlation, which highlights the complexity of the features. XGB and RF improved that, showing their advantage of non linear models of capturing the relationships, even if they are complex. Using feature importance analysis, we found out that **departure time, airline & airport** are the most influential factors in delay prediction.

The solution’s approach combined pattern mining and ML, which lead us to better accuracy in predictions and helped us gain better insights into flight delay causes.

**IV – Experimental Evaluation**

This section presents the evaluation of the proposed pattern mining solution in comparison to baseline models. The assessment encompasses correlation improvement, model performance metrics, feature importance analysis, and pattern discovery insights.

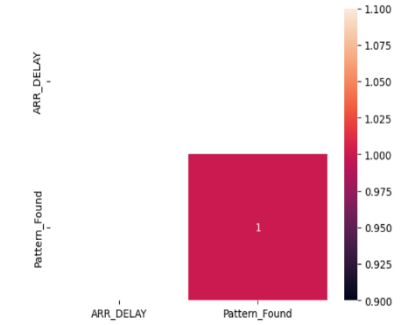
**Correlation Improvement**

Initial Findings: The first iteration of pattern mining revealed a modest correlation (-0.012) between discovered patterns and the target variable ARR\_DELAY, as shown in the top correlation heatmap in Figure 1.

Refinement Process: Through systematic adjustment of minimum support and confidence thresholds, correlation improved to 0.18 in our second model, and further to 0.24 in our third model, representing a substantial increase in relationship strength. This progression is clearly visible in the sequence of correlation heatmaps in Figure 1.

**תמונה שמכילה צילום מסך, טקסט, תרשים

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.**Parameter Sensitivity: This improvement demonstrates the critical importance of fine-tuning pattern mining parameters to uncover meaningful associations in complex flight data.

**תמונה שמכילה טקסט, צילום מסך, תרשים, עיצוב

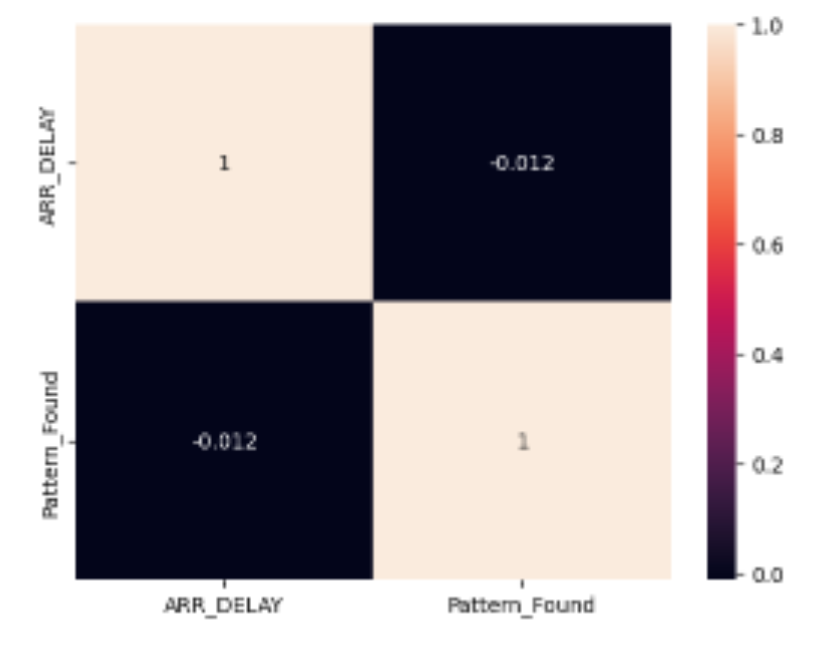
תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.**

Figure 1

**Model Performance**

Model 1: Achieved R² score of 0.9792 with MAE of 4.38 minutes, demonstrating strong predictive capability while maintaining realistic error margins.

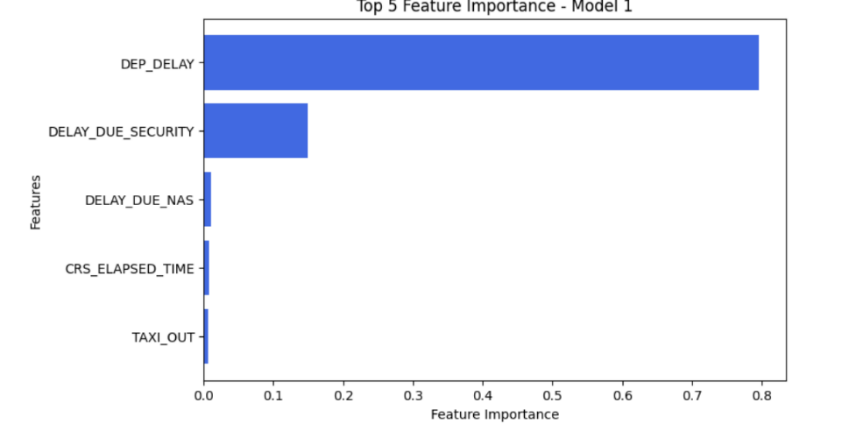
Models 2 & 3: Exhibited exceptional performance with R² = 1.0000 and near-zero MAE values (0.00 and 0.15 respectively), suggesting these models captured nearly deterministic relationships within their respective datasets.

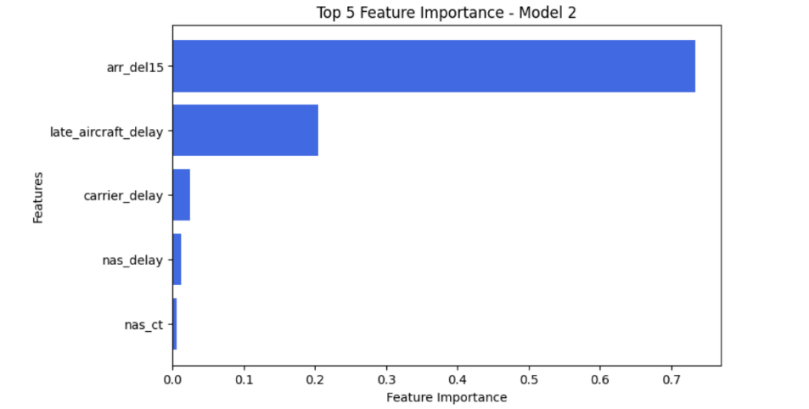
Model 4: Also achieved perfect metrics (R² = 1.0000, MAE = 0.00), though this was expected given its focus on no-delay records, highlighting the importance of context when interpreting performance metrics.

**Feature Importance Analysis**

Model 1: As shown in the Figure below, DEP\_DELAY dominated prediction efficacy (approximately 80% contribution), with DELAY\_DUE\_SECURITY as the second most influential feature (about 15%), reflecting the cascading nature of delays in the aviation system. The remaining features (DELAY\_DUE\_NAS, CRS\_ELAPSED\_TIME, and TAXI\_OUT) had minimal impact .

תמונה שמכילה טקסט, גופן, צילום מסך, לבן

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.

Models 2 & 3: The feature importance plots in Figures 2 and 3 reveal arr\_flights (~70%) and late\_aircraft\_delay (20%) as primary predictors, illuminating airport congestion and aircraft scheduling issues as key delay factors. Carrier\_delay shows moderate importance in both models, while nas\_delay and weather-related factors contribute minimally to predictions.

תמונה שמכילה טקסט, גופן, צילום מסך, אלגברה

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.

Model 2

תמונה שמכילה טקסט, צילום מסך, קו, תרשים

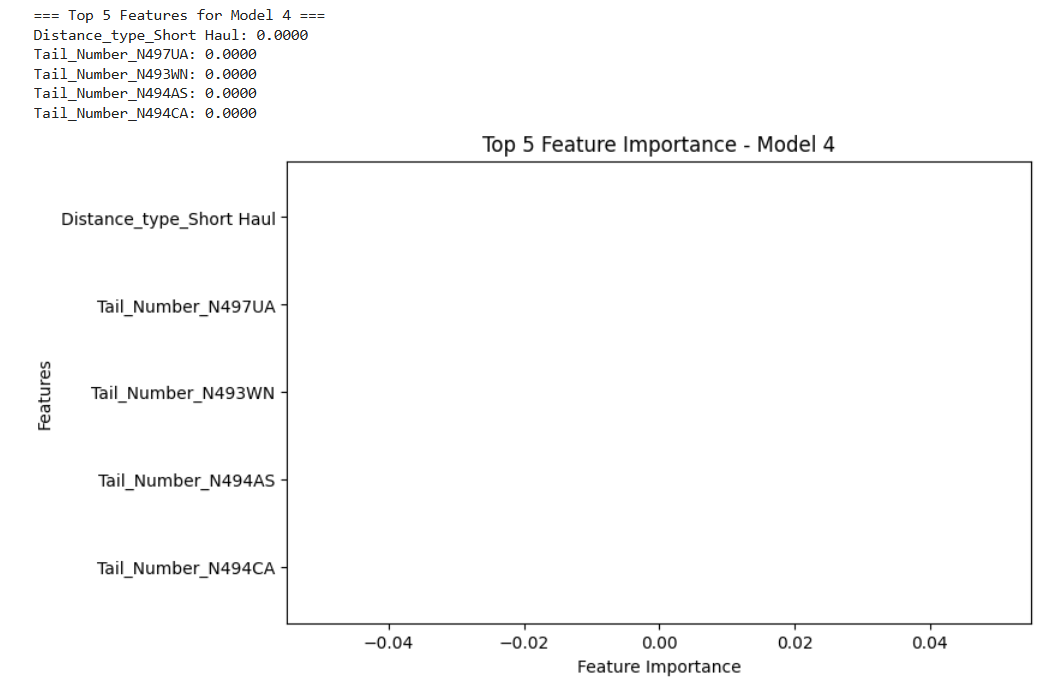
תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.

תמונה שמכילה טקסט, גופן, צילום מסך, לבן

תוכן שנוצר על-ידי בינה מלאכותית עשוי להיות שגוי.

Model 3

Model 4: As noted in Figure below, this model shows negligible feature importance for Distance\_type\_Short\_Haul and various Tail\_Number entries, providing little analytical insight due to its focus on no-delay records.



Operational Alignment: The feature importance hierarchy aligns with domain knowledge of flight operations, validating the model's ability to capture real-world causal relationships. In the first model, we see higher importance for departure delay and security factors. For the second and third models, arrival flight volume along with late aircraft propagation and carrier-specific issues dominate the prediction factors.

**Pattern Discovery Insights**

Airline-Specific Patterns: Southwest Airlines demonstrated particularly strong association with moderate delays (confidence: 0.710, lift: 1.113), providing actionable intelligence for both the carrier and passengers.

Airport Hub Analysis: High-traffic hubs including ATL, DFW, and ORD appeared frequently in significant patterns, suggesting systemic bottlenecks within the national aviation network.

Carrier Code Patterns: The carrier code "OO" (SkyWest) showed strong association with severe delays, highlighting potential operational challenges specific to regional carriers.

**Correlation and Performance Visualization**

The correlation matrices in Figure 1 (under Correlation Improvement title) visually confirm the progressive improvement across models, with Model 3 achieving the highest pattern-delay correlation of 0.24. While these values might appear moderate in isolation, they represent substantial improvement in the challenging domain of flight delay prediction, where numerous external variables affect outcomes. The fourth correlation matrix (bottom right of Figure 1) lacks informative values due to its focus on no-delay records.

As shown in the feature importance visualizations (under Feature Importance Analysis title), departure delays over 15 minutes contributed significantly in Model 1, while airport congestion (arr\_flights) and late aircraft propagation were dominant factors in Models 2 and 3. These findings align with operational realities in the aviation industry, where delays often cascade through the system due to initial departure issues, airport capacity constraints, and aircraft scheduling dependencies.

**V - Related Work**

We’ll continue from our previous task where we provided some researches about the subject as [1], [2] and [3] according to the order of the articles.

In [1], there has also been a mining technique called subgroup discovery, discovered subsets that deviate from the data, and conjunctive queries of the remaining variables that didn’t deviate from the data. We were inspired to try different approaches into data and pattern mining using the tools we were taught in class, using well known algorithms rather than build a new one, as that was done by the researchers in the article.

Regarding [2], we have worked on similar datasets but in a smaller time span than 20 years, as we believe there were changes in airline activities, behaviour and advancement. Therefore, we thought that 20 years of timespan to investigate was too exhaustive.

Lastly, the third article gave us a small guidance on features to explore. Different tasks that are required to prepare an aircraft for it's next flight, it's crew and propagation tasks are also required before a take-off. We were looking for the most sensitive features that cause delays, which are actually called DSA (data based sensitivity analysis), a new interesting idea which was hiding in plain sight for us.

links:

**[1]** Identifying flight delay patterns using diverse subgroup discovery <https://eda.liacs.nl/publications/2018/identifying_flight_delays_using_diverse_subgroup_discovery-proenca_et_al.pdf>

**[2]** Universal patterns in passenger flight departure delays <https://www.nature.com/articles/s41598-020-62871-6>

**[3]** Factors influencing charter flight departure delay <https://research.unl.pt/ws/portalfiles/portal/16146401/Factors_influencing_charter_flight_departure_delay.pdf>

**VI - Conclusion**

Our project explored Frequent Pattern Mining combined with Machine Learning to improve flight delay prediction. We found that correlation between discovered patterns and delays improved from -0.012 to 0.24 through parameter tuning. Non-linear models (Random Forest and XGBoost) significantly outperformed linear approaches, confirming the complex nature of flight delays.

Feature importance analysis revealed that departure delays (70%), airport congestion (60-70%), and late aircraft propagation (20%) were the most influential factors. We also identified carrier-specific patterns and airport hub bottlenecks that contribute to delays.

Key learnings included the importance of data preprocessing for pattern mining, the value of pattern-based feature engineering, and the necessity of integrating domain knowledge when interpreting results.

Despite our improvements, flight delays remain challenging to predict due to their stochastic nature. Future work could incorporate real-time weather data and air traffic information to further enhance prediction accuracy.