Project Background

Meteorites are pieces of space material that survive their journey through Earth's atmosphere and land on the surface. Studying meteorites helps scientists understand the solar system and Earth's history. This project focuses on analyzing a large dataset from NASA, which includes detailed information about over 45,000 meteorite landings, such as their names, types, sizes, dates, and locations.

The goal of the project is to uncover patterns in meteorite landings. First, we aimed to find where meteorites most commonly fall. Second, we explored trends in their types and sizes over time. Finally, we aimed to predict which areas might be at higher risk for future meteorite impacts. To achieve this, we use maps, data mining techniques, and prediction models. We also looked for connections between a meteorite's size, type, and the year it fell.

Methods

We began our analysis by building a geospatial visualization that could explore the distribution of meteorite landing. Using the NASA Meteorite Landings dataset, we used the columns latitude and longitude to plot each meteorite that NASA had collected information on. The dataset was cleaned and missing data in the coordinate column were removed. Meteorite locations were grouped to identify patterns in specific regions. We plotted both a cluster map and a heat map in order to see the exact drop points of the meteorites but also an overall picture with a heatmap. These visualizations were made interactive to allow users to zoom in on hotspots and explore individual meteorite entries.

The cluster map was made using the Folium library in Python and Markercluster() which created the clusters of meteorite landings based on proximity. The heatmap was made also using Folium and a plugin HeatMap() which helped place the color on the map based on the density of meteorites in the area.

After the geospatial visualization, we conducted an association-rule mining analysis where we aimed to identify patterns and co-occurrence relationships within the dataset. We preprocessed the data by dropping rows with any missing data. The dataset includes key variables such as recclass (meteorite classification), fall (whether a meteorite was "Fell" or "Found"), mass (g), and year of discovery. To simplify the analysis and focus on meaningful trends, two transformations were applied: binning and encoding. Year data was grouped into three categories: Pre-1900, 1900-2000, and Post-2000. Mass (g) was categorized into Small (<1000 g), Medium (1000g-5000g), and Large (>5000g). The data was then converted into a transactional format where each meteorite was represented as a set of attributes (e.g., Large, 1900-2000, Iron). The Apriori algorithm was then applied to discover patterns and relationships. We used the following metrics to assess the strength of rules: support, confidence, and lift.

Since the data was dominated by "Found" meteorites and those discovered in the late 20th century and beyond, making it challenging to find meaningful patterns with high support and lift,

we decided to focus on two subsets of the data: meteorites discovered before 1900 and those categorized as “Fell” (meteorites observed falling). By focusing on these subsets, we might uncover more meaningful patterns.

We also created predictive models to analyze and understand meteorite events, focusing on their likelihood of falling into either of the target variable classes: “Fell” or “Found.” The data was preprocessed for modeling by converting geolocation data (latitude and longitude values) into corresponding country names using reverse geocoding libraries in Python. Entries with unaccounted locations were transformed to ensure data completeness. Variables deemed less insightful, such as geolocation fields, nametype, id, and meteorite names, were removed. The dataset was then split into training, validation, and test sets in a 70:15:15 ratio. The target variable, fall, was converted into a binary format: “Fell” events were labeled as 1, while “found” events were labeled as 0.

We built five different models during the process. The first model was a baseline logistic regression model with the parameter class_weight = ‘balanced’. This parameter helped the model learn patterns in the dataset despite the class imbalance in the target variable. The other models included a K-Nearest Neighbors (KNN) model with k=5, a Random Forest model with n_estimators=100, and a Decision Tree model with max_depth=5. Additionally, an Isolation Forest model was built with n_estimators=100 and contamination=0.025.

The contamination parameter represents the expected percentage of the minority class in the target variable. Isolation Forest is particularly well-suited for detecting anomalies in large datasets like this. It identifies anomalies, rare or unusual data points, by isolating instances with fewer splits in a decision tree. Since "Fell" events are less frequent, the model treats them as anomalies, so it is able to address class imbalance and emphasize the distinct characteristics of meteorite falls for deeper analysis.

Key Observations

Through the geospatial visualization, we learned that there were clear patterns in where meteorites are found. The findings were shaped by both natural conditions and human efforts. Antarctica had the most meteorite recoveries because the icy landscape makes meteorites stand out and easier to collect. Similarly, deserts in Northwestern Africa, like those in Algeria and Libya, and in the Arabian Peninsula, such as Oman, also showed many meteorite discoveries. The dry conditions in these areas help preserve meteorites, making them easier to find. In contrast, places like the United States, Europe, and Australia had moderate numbers of meteorite recoveries. These regions have environments that are somewhat suitable for finding meteorites, but not as ideal as deserts or polar areas. On the other hand, places with dense vegetation, like the Amazon rainforest, had very few or no recorded meteorites. This is likely because thick vegetation and poor accessibility make it hard to detect and recover meteorites.

These findings highlight that where meteorites are found depends not only on natural features, like ice or desert, but also on how much effort people put into looking for them in those regions.

Additionally, the geospatial visualizations led us to the conclusion that there are not necessarily 45,000+ meteorites, but instead, often they have broken into pieces and logged as a new finding. So the hotspots could be a couple of complete meteorites broken into a lot of pieces.

The association rule mining analysis revealed distinct patterns within the underrepresented subsets. For meteorites discovered before 1900, a strong association emerged: if a meteorite was large and classified as “Found,” it was likely to belong to the Iron, IIIAB group. This rule, with a lift of 2.66, indicates a strong relationship between these attributes. Pre-1900 meteorites were predominantly large and iron-rich, likely because their durability and visibility made them easier to preserve and collect over time. This reflects a historical collection bias, where large, obvious, and durable specimens were prioritized.

In the “Fell” subset, a different pattern was observed. Medium-sized meteorites of the L6 classification, which are stony in composition, were frequently observed falling during the 1900-2000 period. This pattern is likely due to advancements in observational techniques during this era, including improved documentation and systematic searches. These findings suggest that, compared to the pre-1900 period, the 20th century saw an increase in the recording of smaller, non-metallic meteorites, reflecting the broadening scope of scientific interest and technological capabilities.

These results highlight how historical and technological contexts shaped meteorite data. In the pre-1900 era, large metallic meteorites were overrepresented due to their durability and the practical constraints of early collectors. By contrast, the 20th century saw the inclusion of smaller, less durable meteorites due to advancements in observation and systematic collection methods. Despite these findings, the overwhelming dominance of “Found” meteorites and the late 20th-century period in the dataset made it challenging to uncover high-support patterns within the underrepresented datasets.

The findings from the predictive models varied in results and performance levels. Our baseline model, the Logistic Regression model, achieved a validation accuracy of 96% and a test accuracy of 95%. The KNN model, with *n_neighbors=5*, exhibited slightly higher accuracies of 99% for validation and 98% for test data. From the feature importance rankings of both models, we identified *year* as the most influential predictor, followed by *country*, *recclass*, and *mass (g)*. Similarly, the Random Forest model (*n_estimators=100*) demonstrated high accuracies of 99% (validation) and 98% (test), while the Decision Tree model (*max_depth=5*) achieved 98% accuracy for both validation and test sets. For these models, *year* ranked as the most critical feature, followed by *country*, *mass (g)*, and *recclass*.

In hindsight, the model performances appeared suspiciously high, indicating potential overfitting. Symptoms of overfitting are evident in the high accuracy levels, with several reasons for why this is concerning. The test and validation datasets closely resemble the training data, suggesting that accuracy may not be the best metric for evaluating these models. Additionally, the models were trained using only four features, which may limit the generalizability and depth of insights. Another significant issue is the severe class imbalance in the target variable, with significantly more instances of “Found” events compared to “Fell” events.

Given these observations, the Logistic Regression model emerged as the most reliable option. It effectively addressed the class imbalance issue in the dataset and demonstrated reasonable accuracy with fewer symptoms of overfitting. The Logistic Regression model revealed that the most influential feature was *year*, with a negative coefficient indicating that more recent meteorite events are less likely to belong to the "Fell" class. This suggests that the likelihood of observing meteorites during their fall has decreased in recent years. The feature *country* also had a minor negative impact, while *recclass* and *mass (g)* contributed minimally. These findings highlight temporal patterns as the primary predictor in this dataset.

To gain further insights into the minority class ("Fell"), we applied the Isolation Forest model for anomaly detection. This approach aimed to identify rare meteorite events and uncover interesting patterns in the historical data. The Isolation Forest model achieved validation and test accuracies of 97%, suggesting its ability to detect rare patterns in the dataset. Its feature importance analysis identified *year* as the primary predictor, followed by *mass (g)*, *country*, and *recclass*.

To further explore the relationship between the target variable and the two most important features (*year* and *mass (g)*), we created bar graphs. For the *year* variable, we calculated the proportion of anomalies relative to the total number of events, filtering for the most anomalous years—those with the highest proportion of "Fell" events compared to total events. The bar graph revealed that the proportion of "Fell" events decreased over time, indicating that the likelihood of observing meteorites during their fall has diminished in recent years. For the *mass (g)* variable, we categorized the data into intervals for easier visualization and filtered the top anomalous events within each category. The visualization showed that heavier meteorites are more likely to fall and be observed, while lighter meteorites are less likely to be observed during their fall.

Conclusion/Next-steps

This project has provided valuable insights into the patterns and factors influencing meteorite landings, drawing on geospatial visualization, association rule mining, and predictive modeling techniques. Our analysis reveals that meteorite discoveries are heavily influenced by environmental conditions and human activity, with regions like Antarctica and deserts in North Africa standing out due to their favorable preservation and collection conditions. Additionally, historical biases were evident, as earlier collections predominantly included large, metallic meteorites, whereas the 20th century saw an increase in the collection of smaller, more diverse specimens due to technological advancements. Predictive modeling further underscored the influence of temporal and geographic factors, with the year of discovery being the most critical predictor for meteorite events.

Despite these findings, we faced several challenges throughout the completion of the project. The dataset's dominance by the "Found" meteorites and the inherent class imbalance posed significant obstacles for uncovering patterns in underrepresented categories, such as "Fell" meteorites. Additionally, while the predictive models achieved high accuracy, signs of overfitting suggest that further refinement is needed to ensure generalizability. The Isolation Forest model

showed promise for detecting anomalies and rare events, but its reliance on limited features may constrain its explanatory power.

There are several ways to improve this project in the future. One important step is to handle the imbalance in the dataset between “Fell” and “Found” meteorites. Techniques like oversampling (adding more examples of “Fell” meteorites) or undersampling (removing some “Found” examples) could help balance the data and make the models more reliable. Another idea is to use clustering methods to group meteorites based on their features. Algorithms like k-means or hierarchical clustering could help us understand how different types of meteorites are similar or different, which could improve how we classify them. Finally, adding more information to the dataset could make the models better. For example, including data about the environment where meteorites were found or how much effort was spent looking for them could reveal new patterns and fix some of the overfitting issues we noticed. By exploring these ideas, future work can improve the accuracy of the models and help us learn more about meteorites and their patterns.

Individual contributions

[redacted] - I contributed by creating the cluster geospatial visualizations and the heatmap and analyzing the findings from both maps. I also contributed to the presentation and report.

[redacted] - Contributed to the association rule mining, with the final presentation and explained the findings. Also contributed to the final report.

Jayapreethi Radhakrishnan Madanraj - I contributed by building predictive models and assisting with the final presentation and elaborating the process and findings in the final report.