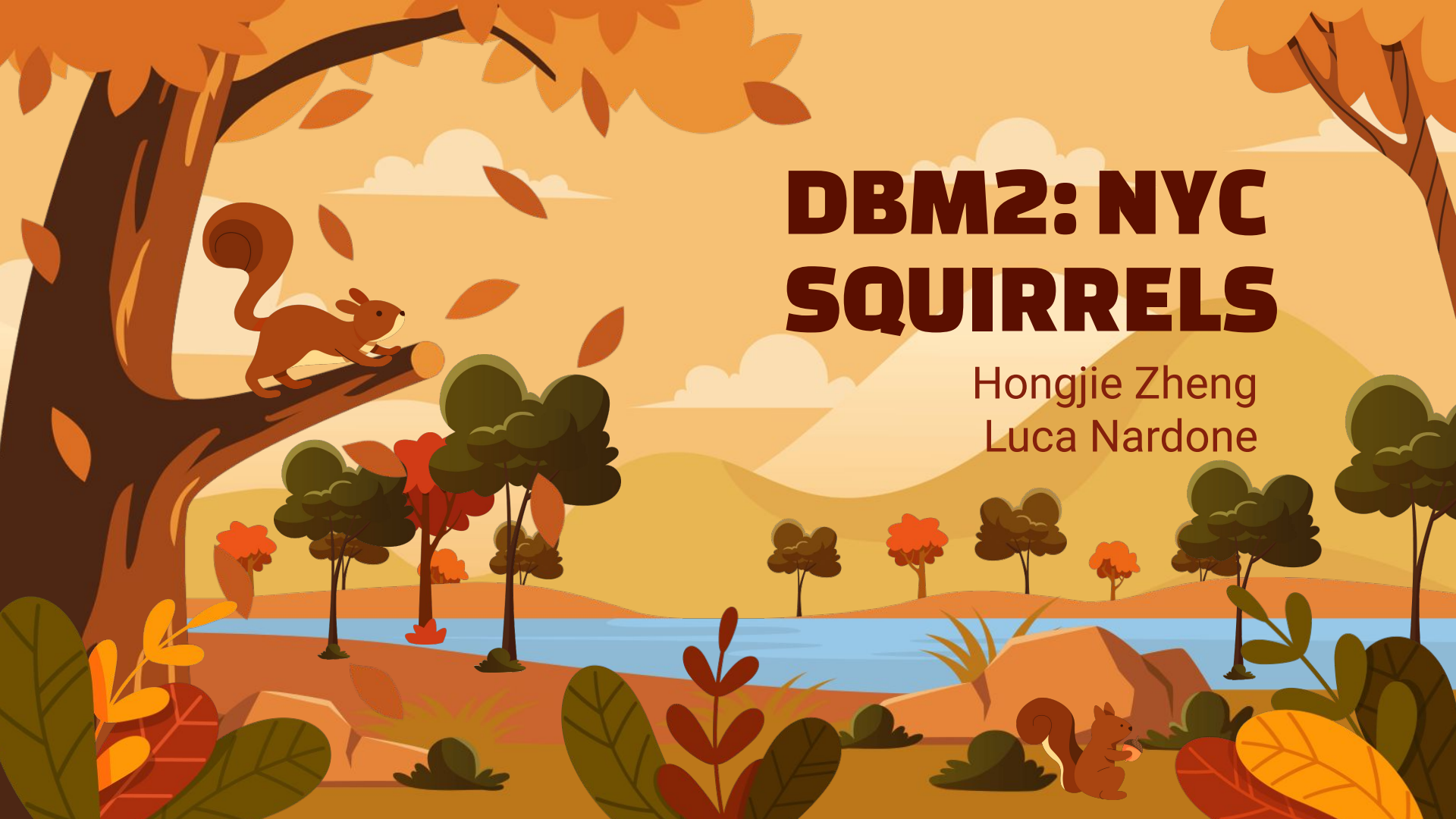


DBM2: NYC SQUIRRELS

Hongjie Zheng
Luca Nardone

Dataset and Preprocessing

🐿️ ***NYC Squirrel Census dataset***→ 3,023 sightings with 36 attributes

Removing duplicates

This ensures that each row in the dataset corresponds to a unique squirrel

Checking for missing values

Identify any missing values in the dataset, which might need to be handled before proceeding

Dropping irrelevant columns

Dropping irrelevant columns, with too much missing data, that won't contribute to the analysis

Handling missing data

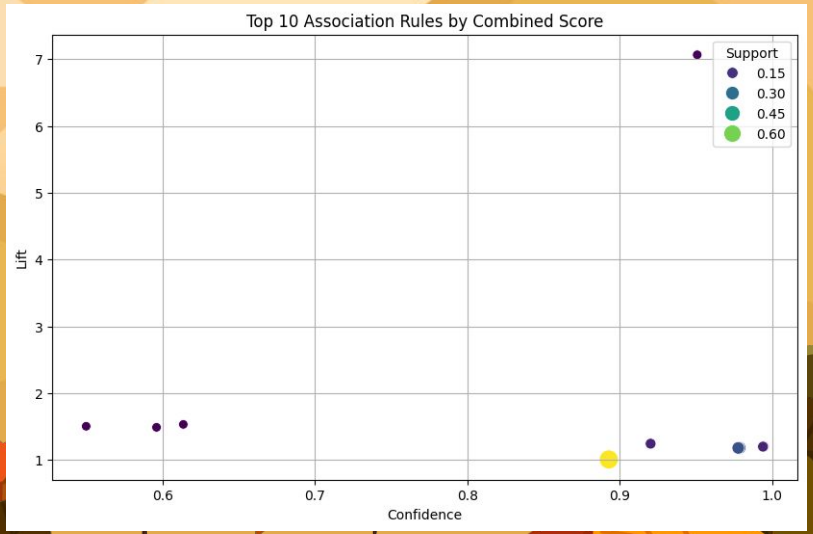
By replacing some of them with the string Unknown and discarding rows with missing values in critical columns

Frequent Pattern Mining

Identify frequent patterns and associations between fur color and behaviors.

1. **ONE-HOT** encoding to converts data into binary format
2. **Apriori algorithm** to find frequent itemsets with minimum support of 5%
3. Generate **association rules** based on the frequent itemsets and filter them to retain only the strongest rules
4. Create a **scatter plot** to visualize the top 10 association rules

Rule	Support	Confidence	Lift
(Running + Cinnamon Fur → Gray Fur)	5.97%	97.64%	1.18
(Running + Unknown Fur Color → Gray Fur)	7.66%	84.86%	1.02
(Chasing → Gray Fur)	7.84%	86.17%	1.04
(Foraging + White Fur → Gray Fur)	7.02%	8.45%	**7.07**

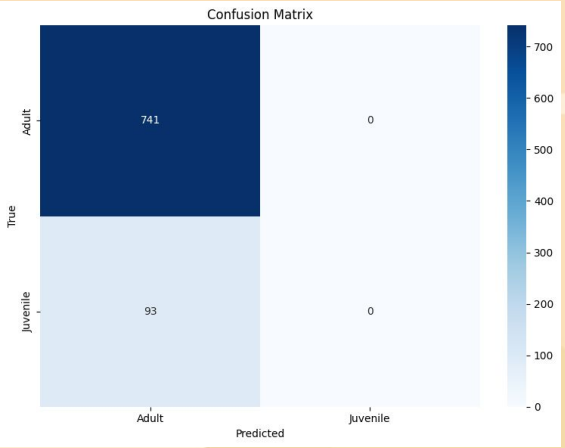


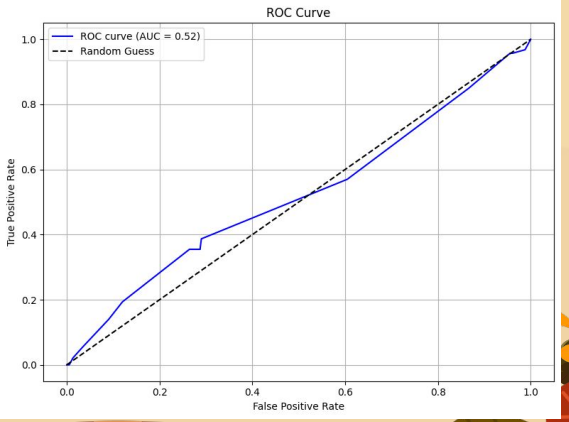
Classification

Classify squirrels into different age groups based on their fur color and behavioral traits

1. **Label Encoding**: Converted categorical variables into numeric values to prepare for classification
2. **Boolean Encoding**: Converted True/False values into binary (0/1)
3. **Train-Test Split**: The dataset was split into training and testing sets, with 30% of the data used for testing and 70% for training
4. **Decision Tree Classifier**: A Decision Tree Classifier was used with the entropy criterion to classify the squirrels into age categories

⚠️ **Class Imbalance problem**: The age categories were highly imbalanced, with a larger proportion of adult squirrels in the dataset.



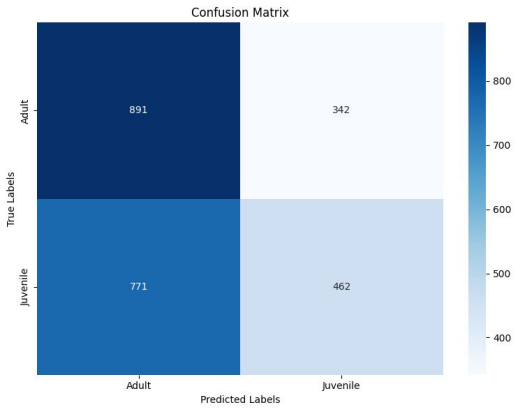


Classification improvements

Balancing the dataset:

- Applied **Random OverSampling** method to duplicate samples from the minority class until the classes are balanced.

The model would learn better from both the minority and majority classes:

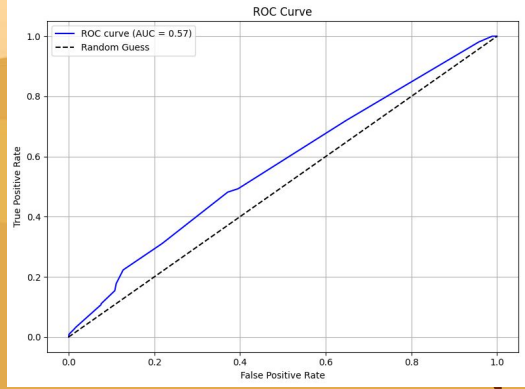


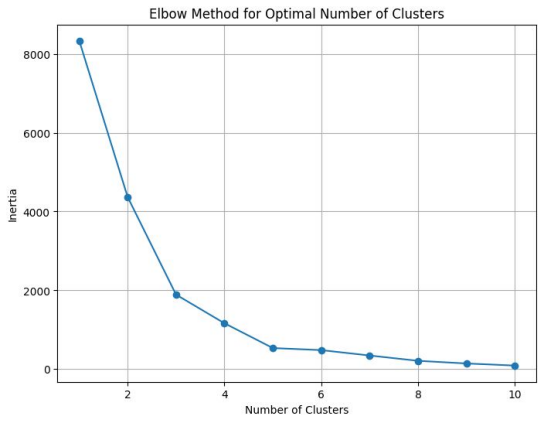
Try to improving Model Results:

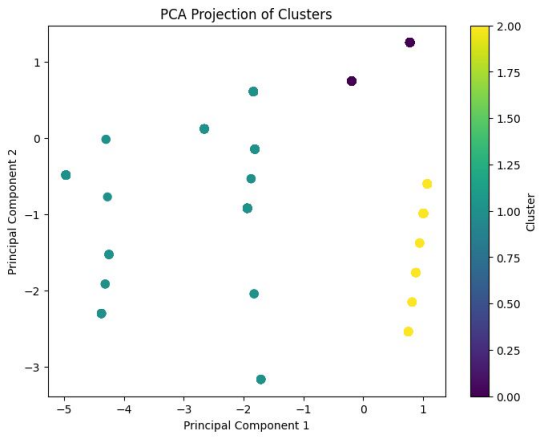
- **Random Forest Classifier** was used to replace the **Decision Tree** model

  It is more robust with his ensemble of trees

This model addressed the imbalance better by aggregating multiple decision trees, each considering different subsets of data:







Clustering

Identify natural groupings of squirrels based on fur colors:

- **K-Means Clustering**
  - Applied the **K-Means** algorithm to identify groups in the data
  - Chose **3 clusters** as optimal using the **Elbow Method**

- **PCA for Visualization**
  - Used **Principal Component Analysis (PCA)** to reduce data dimensions to 2 for easier visualization

*The **Silhouette Score** of **0.71** → clusters are well-defined with minimal overlap.*

Thanks for your attention

For Question:

luca.nardone@insa-lyon.fr
hongjie.zheng@insa-lyon.fr