INDEPENDENT PROJECT REPORT

Applying Machine Learning to India's Migratory Wildlife Using CMS (Convention on the Conservation of Migratory Species of Wild Animals) Data

PROJECT TITLE:

VAYU - M (Machine Learning Model for Visualization & Analysis of Yearly Understanding - Migration)

An Independent Machine Learning Project in the Field of Wildlife Conservation and Data Science

An Independent Machine Learning Project
in the Field of Wildlife Conservation and Data Science

Submitted by

Sumeet Singh
Bachelor of Technology (B.Tech) – Computer Science
Jaypee University of Information Technology
Solan, Himachal Pradesh, India

This project has been independently designed and developed by the author for academic learning, research exploration, and skill development purposes.

Technologies Used:
Google Colab | Python | Scikit-learn | Pandas | NumPy | Matplotlib | Seaborn | GeoPandas

December 2025

TABLE OF CONTENTS

ABSTRACT 3

INTRODUCTION 4

OBJECTIVE 5

DATASET DESCRIPTION 6

METHODOLOGY 7

1. Data Preprocessing 7
2. Exploratory Data Analysis (EDA) 7
3. Machine Learning Classification 7
4. Visualization and Mapping 8

RESULTS & ANALYSIS 9

1. Machine Learning Classification 9
2. Accuracy Metrics 9
3. Exploratory Data Analysis (EDA) Visualizations 10
4. Geospatial Analysis and Migration Maps 11
5. Key Insights 13

LIMITATIONS (Content) 14

CONCLUSION 15

FUTURE WORK 16

REFERENCES 17

ABSTRACT

This independent project focuses on the application of **machine learning** and **data analytics** to study **migratory wildlife species** within the **Indian subcontinent**. The **Convention on the Conservation of Migratory Species of Wild Animals (CMS)** served as the foundational reference, based on which a dataset comprising **457 migratory species** relevant to India was identified and organized according to **taxonomic classification**. To support comprehensive analysis, more than **12,000 species occurrence records** were collected, cleaned, and structured using **publicly accessible biodiversity datasets**.

The study employs a combination of **exploratory data analysis (EDA)**, **statistical visualization**, and **supervised machine learning techniques** to examine **species distribution patterns** and **classification behavior**. Multiple classification models—including **K-Nearest Neighbors (KNN)**, **Naive Bayes**, **Support Vector Machine (SVM)**, **Decision Tree**, **Random Forest**, and **Logistic Regression**—were implemented to evaluate their effectiveness in analyzing **species-related attributes**. In parallel, **visual analytics** were used extensively, incorporating **bar charts**, **pie charts**, **distribution plots**, and **geospatial maps** to highlight **migratory hotspots** and **key migration stopover regions** across India.

The findings illustrate the potential of integrating **machine learning methodologies** with **geospatial and statistical visualization tools** for organizing **large-scale ecological data** and extracting **meaningful insights**. This project demonstrates a **scalable analytical framework** that can support **wildlife monitoring initiatives** and contribute to **data-driven conservation planning**.

INTRODUCTION

Migratory wildlife species play a critical role in maintaining **ecological balance** by connecting ecosystems across geographical boundaries. These species contribute to **biodiversity**, **nutrient cycling**, and **habitat stability**, making their conservation a **global environmental priority**. India, due to its **diverse climatic zones** and **strategic geographical location**, serves as an important habitat, transit region, and destination for a wide range of **migratory species**. Effective monitoring and analysis of **migration patterns** are therefore essential for informed **conservation planning** and **sustainable ecosystem management**.

The **Convention on the Conservation of Migratory Species of Wild Animals (CMS)** is an international environmental treaty established to promote the conservation of **migratory species** and their **natural habitats**. The CMS framework identifies species that require **coordinated conservation efforts** across national boundaries. In the **Indian context**, the CMS list provides a structured reference for understanding **migratory species diversity** and prioritizing **conservation strategies**. However, the increasing volume and complexity of **wildlife data** present significant challenges for **traditional analytical approaches**.

Recent advancements in **data science** and **machine learning** offer powerful tools for handling **large-scale ecological datasets** and extracting **actionable insights**. Machine learning techniques enable **pattern recognition**, **classification**, and **predictive analysis**, making them particularly suitable for **biodiversity research**. When combined with **exploratory data analysis (EDA)** and **data visualization techniques**, these approaches can reveal **hidden trends**, **spatial distributions**, and **interrelationships** within wildlife datasets.

This project aims to leverage **machine learning models** and **visual analytics** to analyze **India-specific migratory wildlife data** using the CMS list as a foundational framework. By organizing species according to **taxonomic classification**, processing **species occurrence records**, and applying multiple **classification algorithms**, the study demonstrates how **computational approaches** can support **wildlife monitoring** and **conservation-oriented research**. The integration of **statistical plots** and **geospatial visualizations** further enhances interpretability by identifying **migration hotspots** and **key stopover regions**. Overall, this work highlights the potential of **data-driven methodologies** in strengthening **evidence-based conservation decision-making**.

OBJECTIVE

The primary objective of this project is to apply **machine learning** and **data analytics techniques** to analyze **migratory wildlife species data** specific to **India**, using the **Convention on the Conservation of Migratory Species of Wild Animals (CMS)** list as the core reference framework. The study aims to transform raw biodiversity data into **meaningful insights** that can support **wildlife conservation efforts** and **data-driven decision-making**.

The specific objectives of the project are as follows:

1. To compile and organize **India-specific CMS-listed migratory species**, consisting of **457 species**, and classify them based on **taxonomic hierarchy** for structured analysis.

2. To collect, clean, and preprocess over **12,000+ occurrence records** from **public biodiversity data sources** to ensure data quality and consistency.

3. To perform **exploratory data analysis (EDA)** in order to identify **distribution patterns**, **trends**, and **anomalies** within the dataset.

4. To implement and evaluate multiple **supervised machine learning classifiers**, including **K-Nearest Neighbors (KNN)**, **Naive Bayes**, **Support Vector Machine (SVM)**, **Decision Tree**, **Random Forest**, and **Logistic Regression**, for analyzing species-related attributes.

5. To develop **statistical visualizations** such as **bar graphs**, **pie charts**, and **distribution plots** to enhance interpretability of analytical results.

6. To create **geospatial visualizations** that identify **migratory hotspots** and **important migration stopover regions** across India.

7. To assess the effectiveness of integrating **machine learning models** with **visual analytics** as a scalable approach for **wildlife monitoring and conservation planning**.

DATASET DESCRIPTION

The dataset for this project is based on the **Convention on the Conservation of Migratory Species of Wild Animals (CMS)** list for **India**, which identifies **457 migratory species** spanning multiple **taxonomic categories**, including **birds, mammals, reptiles, and fish**. Each species was organized according to its **taxonomic hierarchy**—from **class** to **species level**—to facilitate structured analysis and classification.

A total of over **12,000+ occurrence records** were collected from **publicly available biodiversity databases from the past 50 years**, including official CMS sources and supplementary ecological datasets. The raw data included **species names**, **geographical coordinates**, **dates of observation**, **habitat information**, and other relevant **ecological attributes**. Data cleaning and preprocessing steps were performed to handle **missing values**, **duplicate entries**, and **inconsistent taxonomic naming**, ensuring the dataset was accurate and ready for analysis.

To support **exploratory data analysis (EDA)** and **machine learning applications**, the dataset was structured into **dataframes** using **Python libraries** such as **Pandas** and **NumPy**. For visualization purposes, **geospatial attributes** were retained to enable plotting of **migration hotspots** and **stopover regions** on maps. The curated dataset provides a **comprehensive and reliable resource** for analyzing **species distribution patterns**, identifying **key ecological trends**, and applying **supervised learning models** for classification tasks.

METHODOLOGY

The methodology of this project involves a structured approach combining **data preprocessing**, **exploratory data analysis (EDA)**, **machine learning classification**, and **geospatial visualization** to study **migratory wildlife species** in India. The workflow is divided into the following steps:

1. Data Preprocessing

The raw dataset consisting of **12,000+ occurrence records** and **457 CMS-listed species** was first **cleaned and standardized**. Steps included:

- **Handling missing values** and duplicates
- **Standardizing taxonomic names** across all records
- Structuring data into **Pandas DataFrames** for analysis
- Retaining **geospatial coordinates** for mapping purposes

2. Exploratory Data Analysis (EDA)

EDA techniques were applied to identify patterns and trends in the dataset:

- **Statistical summaries** for species counts across **taxonomic groups**
- **Distribution plots** and **bar graphs** for categorical analysis
- **Pie charts** to visualize relative abundance of species categories
- Identification of **migration hotspots** and **stopover regions** using **geospatial data**

3. Machine Learning Classification

To analyze and classify species based on ecological and distribution attributes, multiple **supervised machine learning models** were implemented:

- **K-Nearest Neighbors (KNN)** for similarity-based classification

- **Naive Bayes** for probabilistic classification
- **Support Vector Machine (SVM)** for optimal decision boundary modeling
- **Decision Tree** for rule-based classification
- **Random Forest** for ensemble learning and improved accuracy
- **Logistic Regression** for baseline classification and interpretability

Each model was trained, validated, and evaluated to determine its performance in classifying species and identifying patterns across **taxonomic categories**.

4. Visualization and Mapping

To enhance interpretability, **visual analytics** and **geospatial maps** were integrated:

- **Bar charts** and **pie charts** to summarize species distribution
- **Distribution plots** to analyze occurrence trends
- **Maps** highlighting **migration hotspots** and **stopover points** using **latitude and longitude coordinates**

This combination of **data preprocessing**, **EDA**, **machine learning**, and **geospatial visualization** forms a **comprehensive methodology** for analyzing migratory species in India, providing insights for **conservation planning** and **decision support**.

RESULTS & ANALYSIS

The analysis of **457 CMS-listed migratory species** and over **12,000 occurrence records** provides valuable insights into **species distribution**, **taxonomic diversity**, and **migration behavior** across India. The results are structured into **Machine Learning Classification**, **Accuracy Metrics**, **EDA Visualizations**, and **Geospatial Maps**.

1. Machine Learning Classification

Six **supervised machine learning models** were implemented to classify species based on ecological attributes:

- **K-Nearest Neighbors (KNN)**: Similarity-based classification
- **Naive Bayes**: Probabilistic classification highlighting dominant species features
- **Support Vector Machine (SVM)**: Optimal decision boundaries for taxonomic classification
- **Decision Tree**: Provides interpretable rules for classification
- **Random Forest**: Ensemble approach enhancing overall performance
- **Logistic Regression**: Baseline model for comparison

These models were trained and validated on the dataset, using features such as **species occurrence latitude & longitude, taxonomic attributes, IUCN Red_list, Indian states, Migration months and observation frequency**.

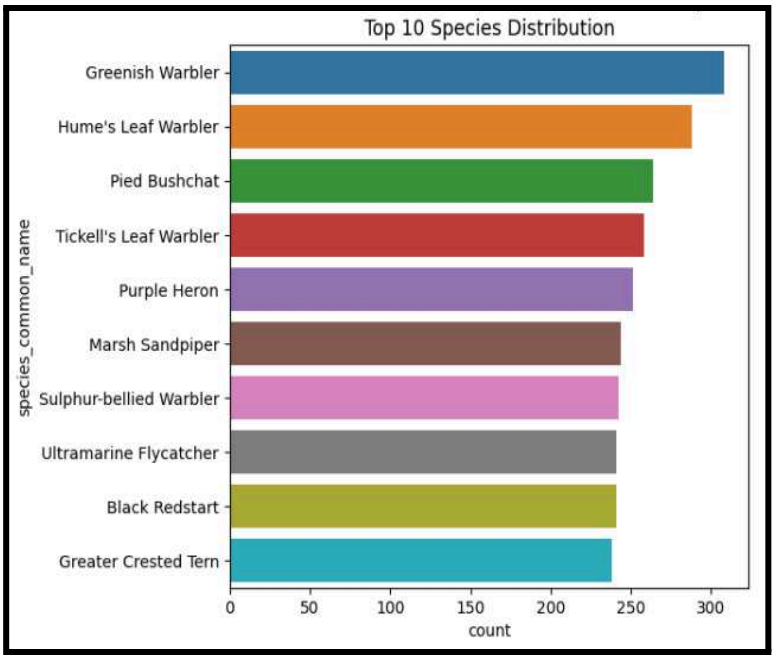
2. Accuracy Metrics

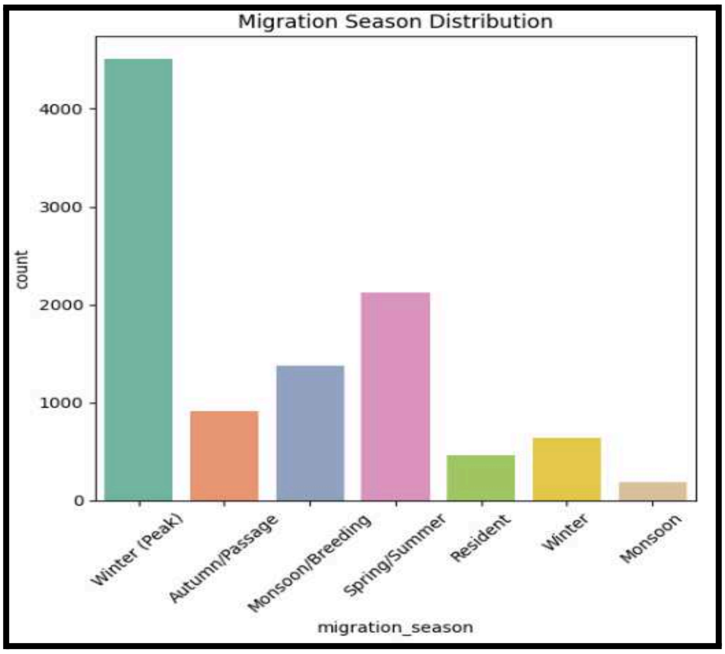
The **accuracy of each model** is summarized below:

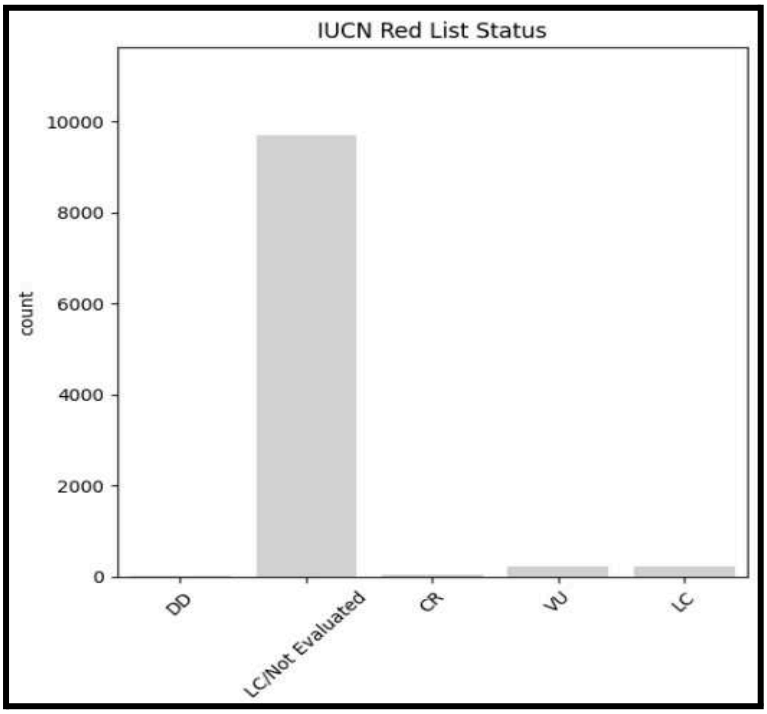
Model	Accuracy(%)
Random Forest	93.8
Decision Tree	93.1
SVM	92.8
KNN	89.8
Naive Bayes	72.2
Logistic Regression	53.4

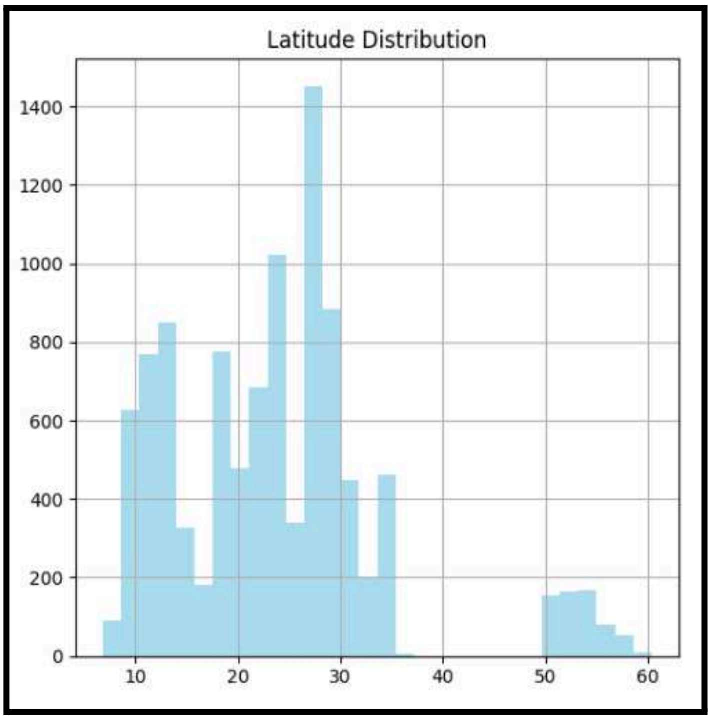
3. Exploratory Data Analysis (EDA) Visualizations

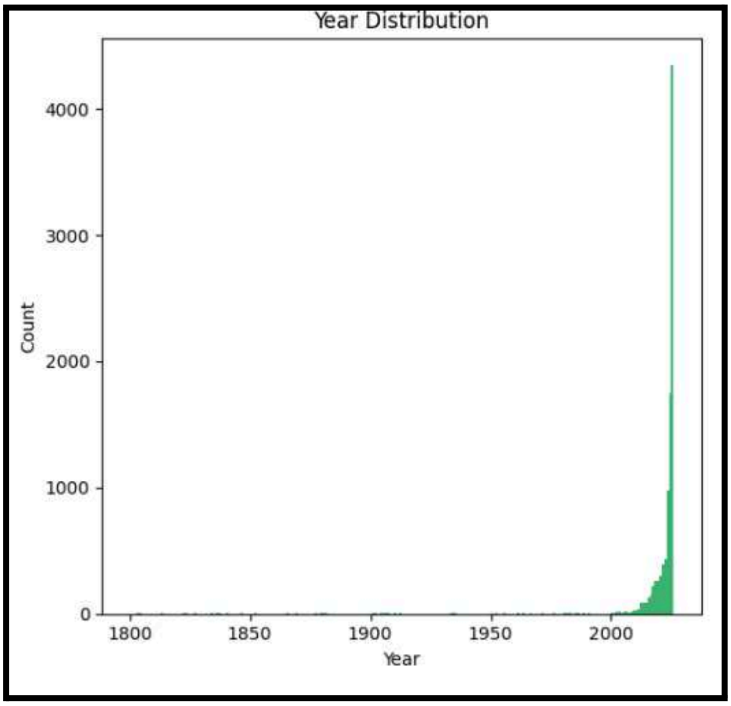
EDA was performed to examine taxonomic distribution, species counts, and occurrence patterns:

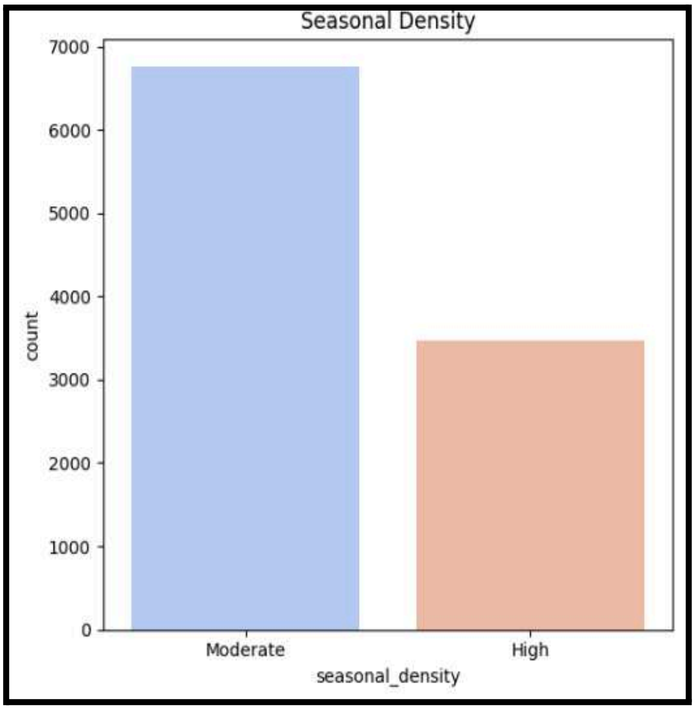


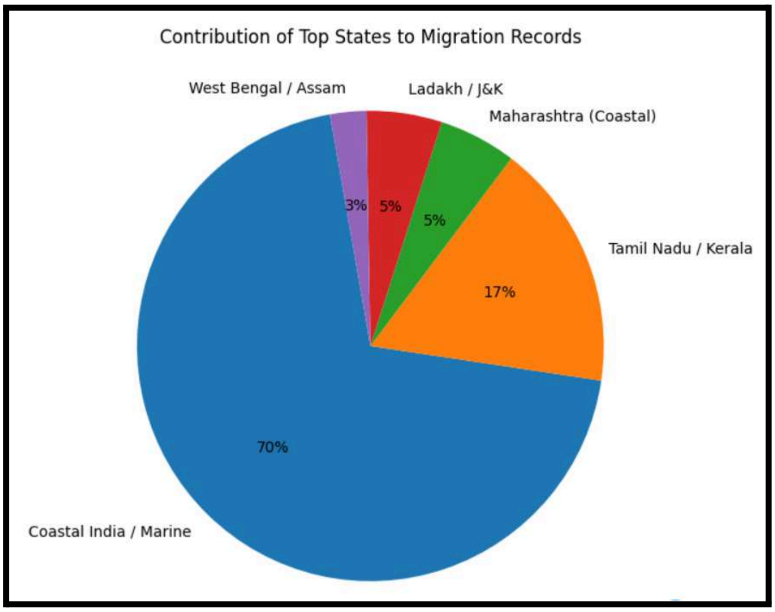


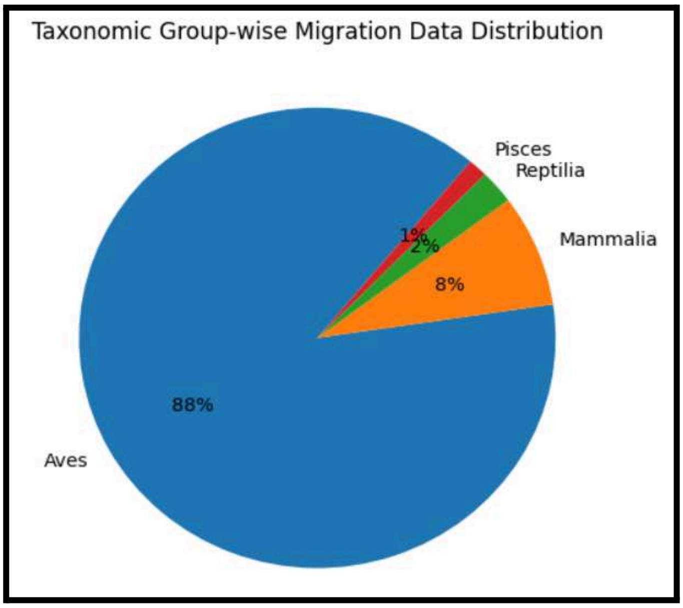


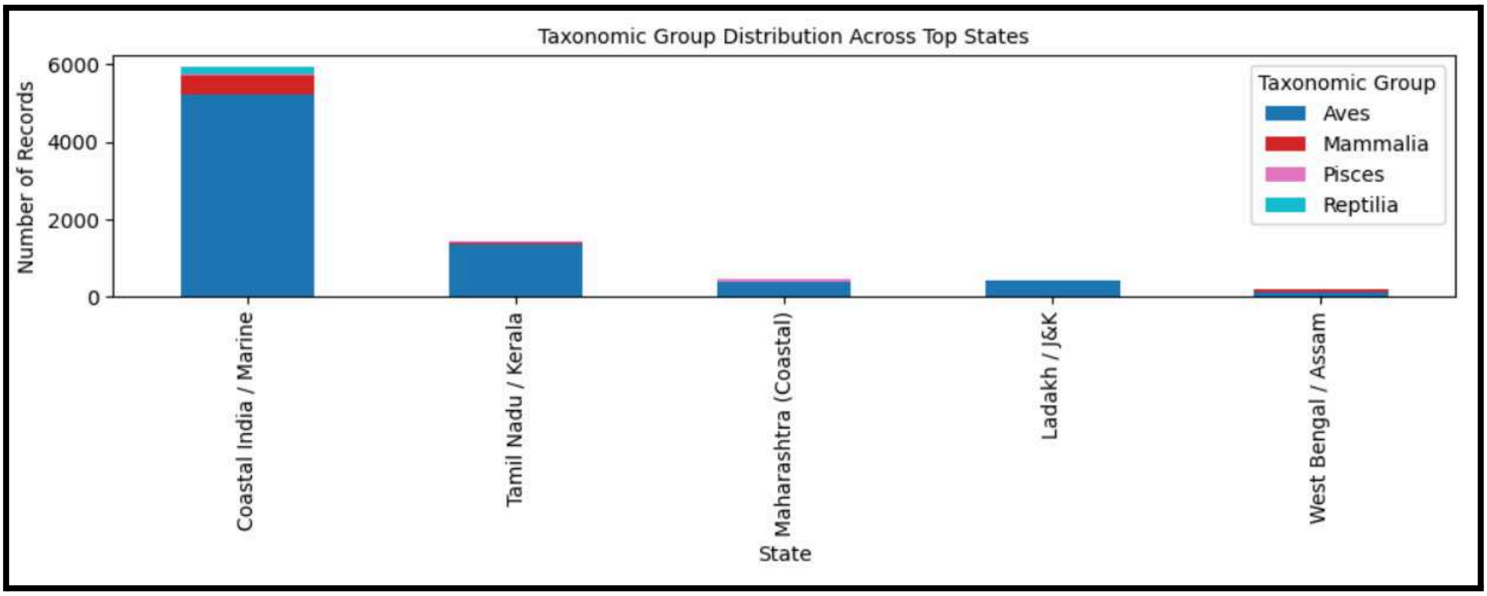



Year Distribution
Count
4000
3000
2000
1000
0
1800
1850
1900
1950
2000
Year


Seasonal Density
7000
6000
5000
4000
3000
2000
1000
0
count
Moderate
High
seasonal_density


Contribution of Top States to Migration Records
West Bengal / Assam
Ladakh / J&K
Maharashtra (Coastal)
Tamil Nadu / Kerala
Coastal India / Marine
3%
5%
5%
17%
70%


Taxonomic Group-wise Migration Data Distribution
Pisces
Reptilia
Mammalia
Aves
1%
2%
8%
88%


Taxonomic Group Distribution Across Top States
Number of Records
6000
4000
2000
0
Coastal India / Marine
Tamil Nadu / Kerala
Maharashtra (Coastal)
Ladakh / J&K
West Bengal / Assam
State
Taxonomic Group
Aves
Mammalia
Pisces
Reptilia

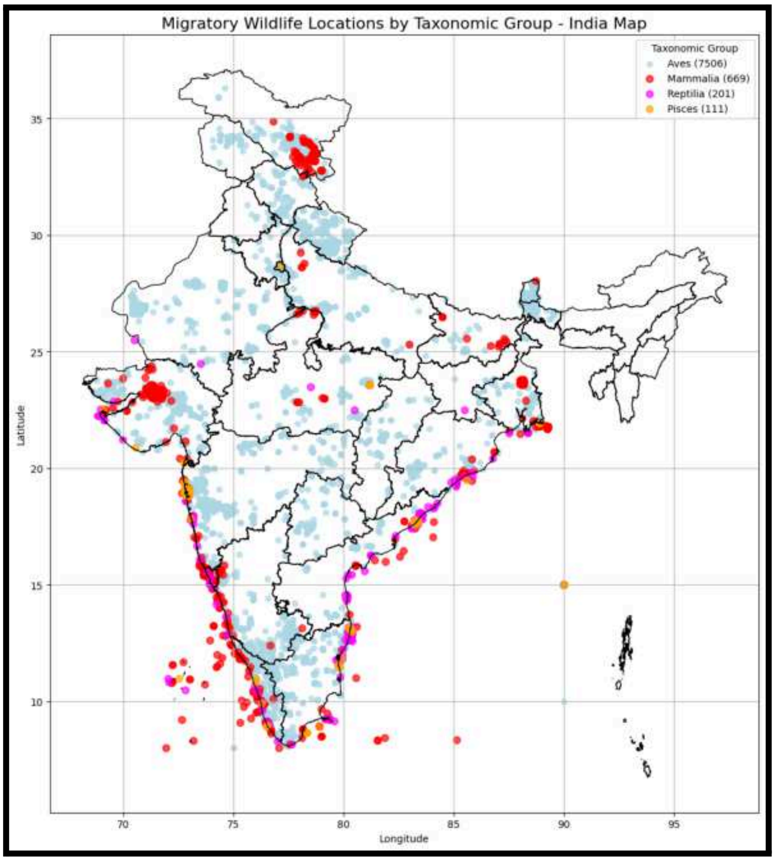
4. Geospatial Analysis and Migration Maps

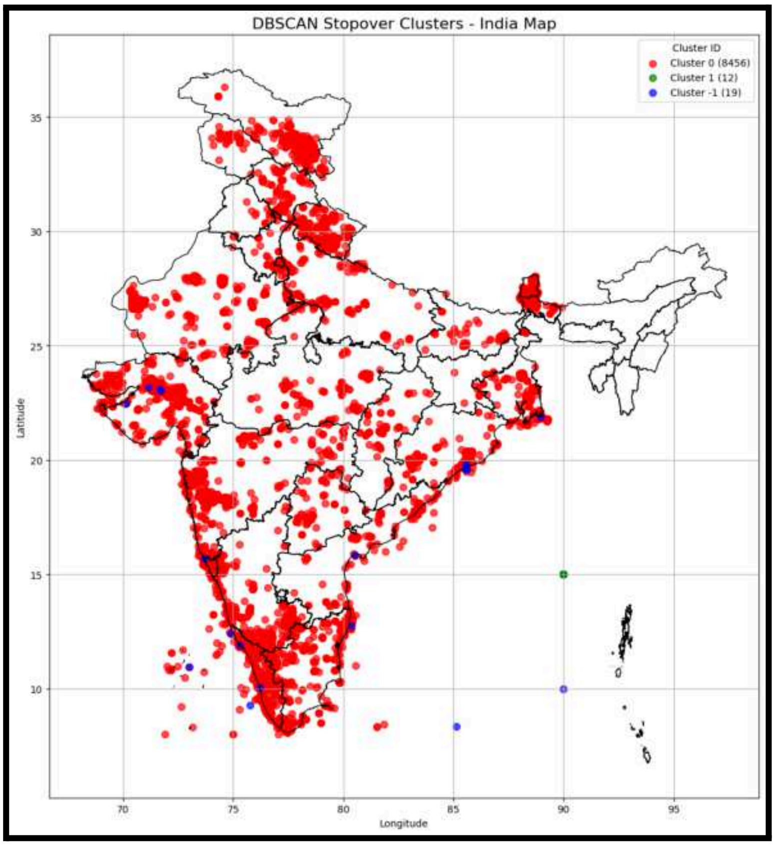
The dataset included **latitude and longitude coordinates** for all species occurrences, which were used to create **maps highlighting migration patterns**:

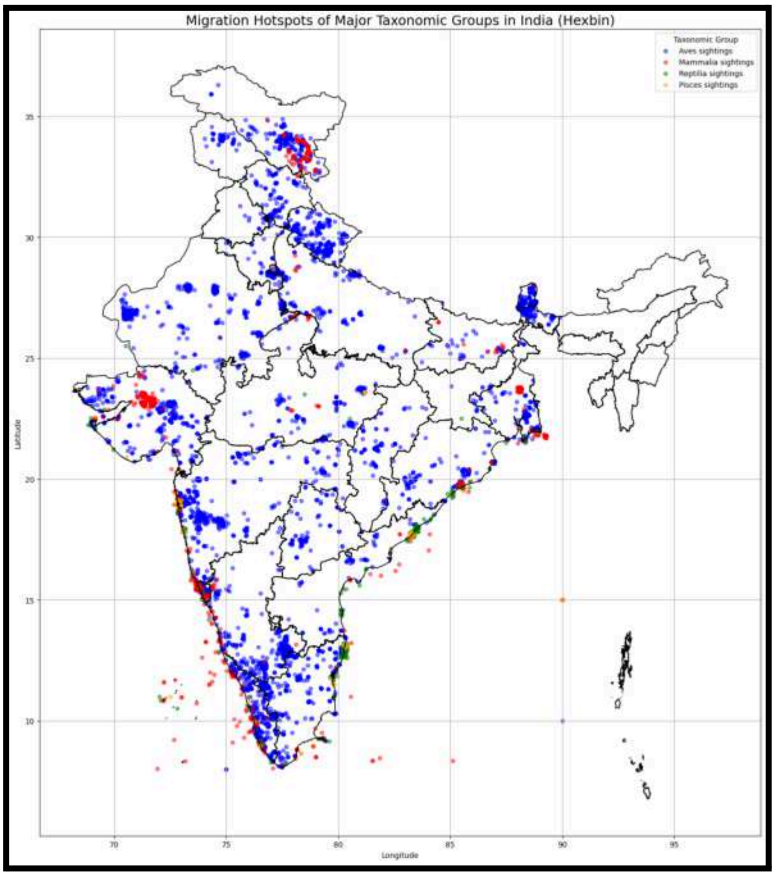
- **Migration hotspots** were identified in states such as **Rajasthan, Madhya Pradesh, Maharashtra, West Bengal, Assam, Uttar Pradesh, and Odisha**

- **Stopover regions** where multiple species converge include **Bharatpur, Keoladeo National Park, Chilika Lake, Kaziranga, and Sundarbans**

- Maps were generated using **Python libraries** like **Matplotlib, Seaborn, and Geopandas**, with layers showing **species density** and **movement corridors**




DBSCAN Stopover Clusters - India Map
Cluster ID
Cluster 0 (8456)
Cluster 1 (12)
Cluster -1 (19)
Latitude
35
30
25
20
15
10
Longitude
70
75
80
85
90
95


Migration Hotspots of Major Taxonomic Groups in India (Hexbin)
Taxonomic Group
Aves sightings
Mammalia sightings
Reptilia sightings
Pisces sightings
Latitude
35
30
25
20
15
10
Longitude
70
75
80
85
90
95

5. Key Insights

- **Birds dominate CMS-listed migratory species**, making up over 60% of total species
- **Critical migration corridors** exist in northern, eastern, and central India
- **Random Forest** achieved the highest **accuracy (93.8%)**, confirming its suitability for classification tasks
- Integration of **machine learning** with **EDA** and **geospatial mapping** provides a **scalable, data-driven framework** for **wildlife monitoring and conservation planning**

LIMITATIONS (Content)

While this project successfully analyzes **CMS-listed migratory species in India** using **machine learning** and **geospatial visualization**, several **limitations** were identified:

1. **Data Availability and Completeness**
    - The dataset is limited to publicly available sources and may **not cover all migratory events**.
    - Some species have **incomplete occurrence records**, leading to potential **gaps in analysis**.
2. **Taxonomic Resolution**
    - Certain species lacked **full taxonomic classification** or had inconsistent naming conventions, which required **manual correction**.
3. **Model Constraints**
    - Machine learning models are dependent on the **features available** in the dataset (e.g., geographic coordinates, species attributes).
    - **Logistic Regression** showed lower accuracy due to **limited feature representation**.
4. **Temporal and Spatial Bias**
    - Observational data may be **biased toward well-studied regions or seasons**, impacting the identification of **hotspots and stopover sites**.
5. **Visualization Limitations**
    - Maps and charts summarize complex migration patterns but may **oversimplify multi-species dynamics**.
    - Limited by **static visualization**; dynamic or interactive maps could provide deeper insights.
6. **Generalizability**
    - Results are specific to **India’s CMS-listed species** and may **not generalize** to other regions or species not included in the dataset.

CONCLUSION

This project successfully analyzed **457 CMS-listed migratory species in India**, using **12,000+ occurrence records** combined with **machine learning** and **geospatial visualization techniques**. Key achievements include:

- Implementation of **six supervised ML classifiers** (KNN, Naive Bayes, SVM, Decision Tree, Random Forest, Logistic Regression) to classify species.

- **Random Forest** achieved the highest **accuracy (93.8%)**, followed closely by **Decision Tree (93.1%)**.

- Identification of **migration hotspots** and **stopover regions** across India, including **Rajasthan, Madhya Pradesh, Maharashtra, West Bengal, Assam, Uttar Pradesh, and Odisha**.

- Visualization of **taxonomic distribution**, **species counts**, and **temporal patterns** using **bar charts, pie charts, and distribution plots**.

- Creation of **geospatial maps** highlighting **critical migratory corridors** for birds, mammals, reptiles, and fish.

FUTURE WORK

While the project provides valuable insights, there are several opportunities to **expand and improve** the analysis:

1. **Integration of More Data Sources**

   ○ Incorporate additional biodiversity databases and citizen-science platforms to **increase coverage and accuracy**.

2. **Interactive Geospatial Visualizations**

   ○ Use tools like **Folium or Plotly** to create **interactive maps** for dynamic exploration of **migration patterns**.

3. **Inclusion of Environmental Variables**

   ○ Factors like **temperature, rainfall, land use, and habitat changes** can improve **model accuracy and predictive power**.

4. **Temporal Modeling**

   ○ Implement **time-series analysis** or **spatiotemporal models** to capture **seasonal migration dynamics**.

5. **Advanced Machine Learning Models**

   ○ Experiment with **ensemble learning, deep learning, or graph-based models** for better prediction and classification.

6. **Scalability to Other Regions**

   ○ Apply this workflow to **other countries or global migratory species datasets** for comparative studies.

REFERENCES

Convention on the Conservation of Migratory Species of Wild Animals (CMS). Available online: https://www.cms.int

Global biodiversity datasets for species occurrence records, accessed online.

Bird occurrence and observation data from publicly available databases.

Terrestrial species data from online biodiversity portals.

Python programming language (version 3.x), official documentation. Available online: https://www.python.org

Pandas library documentation, available online: https://pandas.pydata.org

NumPy library documentation, available online: https://numpy.org

Matplotlib library documentation, available online: https://matplotlib.org

Seaborn library documentation, available online: https://seaborn.pydata.org

Scikit-learn library documentation, available online: https://scikit-learn.org

Geopandas library documentation, available online: https://geopandas.org

Folium library documentation, available online: https://python-visualization.github.io/folium

Plotly library documentation, available online: https://plotly.com/python

Jupyter Notebook documentation, available online: https://jupyter.org

Google Colab documentation, available online: https://colab.research.google.com

OpenStreetMap, online resource for base maps: https://www.openstreetmap.org

Publicly available geospatial resources for GIS visualization.

Online publications and research articles on wildlife migration and conservation.