**pDSC 680 Applied Data Science**

**Chitramoy Mukherjee, T301 T302 2253 Winter 2025**

**Project-2 Milestone-3**

**Wildlife Protection of Endangered Species**

**Description:**

An endangered species is a type of organism that is threatened by extinction. Species become endangered for two main reasons: loss of habitat and loss of genetic variation. The conservation of endangered and rare wildlife is a primary part of preserving biodiversity. The purpose of this project is to intelligently monitor endangered and rare wildlife species by using Machine Learning models with animal data and computational resources from various parks in United States.

**Problem Statement:**

The rapid decline in populations of endangered species globally poses a critical threat to biodiversity and the stability of ecosystems. Traditional methods of monitoring and protecting these species are often labor-intensive, time-consuming, and prone to inaccuracies. There is an urgent need for innovative, scalable, and efficient solutions to identify, monitor, and protect endangered species effectively. Leveraging advancements in Artificial Intelligence (AI) and Machine Learning (ML), we propose developing a sophisticated AI/ML-based system to enhance the efforts of wildlife conservationists, researchers, and policymakers in safeguarding endangered species.

**Objectives:**

1. **Identification:** Develop AI/ML models capable of accurately identifying endangered species in their natural habitats using camera traps, drones, and satellite imagery.
2. **Monitoring:** Implement AI-powered real-time monitoring systems to track the movements, behaviors, and population dynamics of endangered species, providing valuable data for conservation efforts.
3. **Threat Detection:** Utilize machine learning algorithms to predict and identify potential threats to endangered species, such as poaching, habitat loss, and climate change impacts, allowing for timely and targeted interventions.
4. **Data Integration:** Create a centralized platform to integrate and analyze data from various sources, including field observations, sensor networks, and citizen science initiatives, to provide comprehensive insights into the status of endangered species.
5. **Policy Support:** Generate actionable insights and recommendations for policymakers and conservation organizations to formulate and implement effective conservation strategies based on data-driven evidence.

By harnessing the power of AI/ML, this project aims to revolutionize wildlife protection efforts, ensuring the survival and flourishing of endangered species for future generations.

**Datasets and Data Dictionary:**

Here I am using 3 datasets as below:

**File**: final\_species.csv

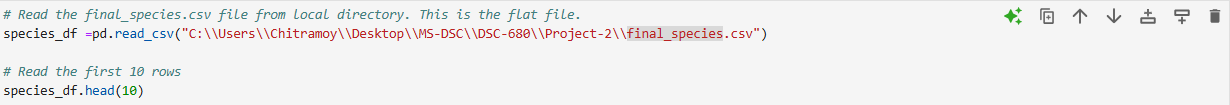
**Source**: Abigail Larion(n.d.), <https://www.kaggle.com/datasets/nationalparkservice/park-biodiversity/data?select=species.csv>

In the final\_species.csv we have added 2 additional derived column is\_float and first\_common\_name for the project purpose.



The species dataset contains plants and animal species information from different national parks of United States. This dataset contains 13 columns.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| Species ID | String | Unique ID. National Park service code for each species |
| Park Name | String | Park Name in which the species appears. |
| Category | String | Category of species like mammal, bird etcetera. |
| Order | String | The scientific order the species belongs to. |
| Family | String | The scientific family the species belongs to. |
| Scientific Name | String | Full scientific species name. |
| Common Names | String | Usual common name(s) for the species. Comma-delimited. |
| Record Status | String | record status from park |
| Occurrence | String | Current presence of that species |
| Nativeness | String | Whether the species is native to the area or a non-native/invasive. |
| Abundance | String | Commonality of sightings. |
| Seasonality | String | When the species can be found in the park.  Blank if the species is found there year-round. |
| Conservation Status | String | IUCN species conservation status. |



**Website**: National Park locations and areas

**Source**: <https://irma.nps.gov/NPSpecies>

Created the final\_web\_species\_data with required columns from park\_name, state, location, date\_established, area, recreation\_visitors, description and established\_date.



This data source will be used to get parks information as below. Further information is provided under “Project Approach” section on how this data will be extracted.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| Park Code | String | Unique value. National Park code |
| Park Name | String | Name of the park. |
| State | String | US state(s) in which the park is located. Comma-separated. |
| Acres | Integer | Size of the park in acres. |
| Latitude | Integer | Latitude of the park (centroid). |
| Longitude | Integer | Longitude of the park (centroid). |

A close up of a computer screen

Description automatically generated

**API**: Animals API

**Source**: [Animals API - API Ninjas (api-ninjas.com)](https://api-ninjas.com/api/animals)

The Animals API provides interesting scientific facts on thousands of different animal species.

/v1/animals

**HTTP GET**: Returns up to 10 results matching the input name parameter.

**Parameters**: name (required) - common name of animal to search. This parameter supports partial matches (e.g. fox will match gray fox and red fox)

**Headers**: X-Api-Key (required) - lJCoL9STz0hLjJXG8gEYpQ==IGRnvx5k0MwcjfXR

**Example API call**:

Using Python script in Jupyter Notebook.

import requests  
name = 'fisher'  
api\_url = 'https://api.api-ninjas.com/v1/animals?name={}'.format(name)  
response = requests.get(api\_url, headers={'X-Api-Key': 'lJCoL9STz0hLjJXG8gEYpQ==IGRnvx5k0MwcjfXR'})  
if response.status\_code == requests.codes.ok:  
 print(response.text)  
else:  
 print("Error:", response.status\_code, response.text)

A screen shot of a computer

Description automatically generated



A screen shot of a computer

Description automatically generated

**Methods/Steps:**

Using the provided sources, here are specific analysis methods to complete the Wildlife Protection of Endangered Species project:

**1. Data Collection and Integration**

* **National Park Locations and Areas (Website):** Scrape or extract data on national park locations, areas, and relevant attributes. This data will provide important contextual information about the habitats and protected areas for endangered species.
* **Animals API:** Access and retrieve animal-related data such as species names, characteristics, habitats, and conservation statuses. This API data can be integrated with other sources to enhance the overall dataset.
* **Species.csv (File):** Import and preprocess the CSV file to extract species-specific data. This file may contain information on species names, scientific classifications, population sizes, threats, and conservation statuses.

**2. Geospatial Analysis**

* **Habitat Mapping:** Using GIS tools, map the locations of national parks and overlay them with species distribution data from the Animals API and Species.csv. This helps in identifying critical habitats and conservation areas.
* **Habitat Suitability Modeling:** Apply machine learning algorithms like MaxEnt to predict suitable habitats for endangered species based on environmental variables (e.g., climate, vegetation) extracted from the National Park data and Animals API.

**3. Species Identification and Monitoring**

* **Species Classification:** Develop and train machine learning models to classify different species based on attributes from the Animals API and Species.csv. Use this to identify species captured in images and videos from camera traps or drones within national parks.
* **Population Monitoring:** Utilize data from the Species.csv file to track population trends over time. Apply time-series analysis to detect patterns and predict future population changes.

**4. Threat Detection and Prediction**

* **Poaching and Illegal Activities:** Implement anomaly detection algorithms on data from the national parks to identify unusual activities or potential threats. Combine this with real-time monitoring using AI-powered surveillance systems.
* **Environmental Changes:** Analyze satellite imagery and remote sensing data to detect habitat changes and deforestation within national parks. Use this data to assess the impact of environmental changes on endangered species.

**5. Data Visualization and Reporting**

* **Interactive Maps:** Create interactive maps to visualize species distributions, habitats, and threats. Integrate data from the National Park locations, Animals API, and Species.csv to provide a comprehensive view.
* **Dashboards:** Develop dashboards to display key metrics and trends related to endangered species protection. Use data visualization tools to present insights and recommendations to conservationists and policymakers.

By utilizing these analysis methods, the project can effectively integrate data from various sources to monitor, protect, and conserve endangered species within national parks and their surrounding areas. If you have any specific requirements or need further details, feel free to let me know!

**Analysis:**

In below steps, create a new SQLLite database named as 'dsc\_680\_final\_prj\_cm.db'. If the database doesn't exist then it will get created otherwise connection will get established for existing database. Once connection gets established, load species\_df into a SQlLite table named 'us\_park\_species'. Similarly, load web\_species\_df into a SQlLite table named 'park\_details' and load api\_species\_df into a SQlLite table named 'species\_details'.

If the table already exists, then it will be replaced. Here table 'us\_park\_species' contains 3711 rows, 'park\_details' table contains 63 rows and 'species\_details' table contains 43 rows.

**A screenshot of a computer

Description automatically generated**

After loading the Dataframes, fetch all the column names from all the 3 tables for better visibility of columns. Here we will use the PRAGMA command to show the structure of columns.

Brief introduction of tables

1. us\_park\_species : Contains information of which US park contains how many types of species and basic details

2. park\_details : Contains US park information like area, lat/lon, totat recreation visitors etc.[¶](http://localhost:8888/notebooks/Desktop/MS-DSC/DSC-680/Project-2/Chitramoy_Mukherjee-DSC-680-Project-2.ipynb?#2.-park_details-:-Contains-US-park-information-like-area,-lat/lon,-totat-recreation-visitors-etc.)

2. species\_details : Contains individual species details like their kingdom, predators, habitat, biggest threar etc.

Create primary indexes on tables 'park\_details' and 'species\_details'. In 'park\_details' table, park\_name column can be primary. In 'species\_details' table, species\_name can be primary.

A white rectangular object with a black border

Description automatically generated

A white background with a black border

Description automatically generated with medium confidence

A white rectangular object with a white background

Description automatically generated

A white screen with red and blue text

Description automatically generated

A screenshot of a computer

Description automatically generated

A white rectangular object with a white background

Description automatically generated

Here we will use a SQL query with the LIKE operator to perform a case-insensitive join on the park\_name columns of both tables 'us\_park\_species' and 'park\_details'. The % wildcard characters are used to match any sequence of characters.

Load the merged data into a dataframe and show top 10 rows.

A close up of a screen

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer error

Description automatically generated

A screenshot of a graph

Description automatically generated

A close-up of a website

Description automatically generated

A screenshot of a computer

Description automatically generated

A close-up of a text

Description automatically generated

A screen shot of a graph

Description automatically generated

**Conclusion:**

The purpose of this project is to intelligently monitor endangered and rare wildlife species by using Machine Learning models with animal data and computational resources from various parks in United States. For this project, data has been collected from flat file, website and API sources. The data extracted from flat file contains plants and animal species information from different national parks of United States. After applying few cleaning and transformation steps, final dataframe was generated and saved locally as final\_species.csv. Next set of data which got extracted from Website contains park information which includes Location, Date established, Area of the park and Number of visitors information of all the national parks in United States. Post cleanup and transformation, this dataset was saved locally as final\_web\_species\_data.csv. The 3rd set of data was extracted from Animal API which contains detail information for individual species like Average lifespan, Biggest Threat, Habitat, Predator etc. This dataset got saved locally as final\_api\_species\_data.csv after cleanup and transformation.  
The purpose of this project is to merge all three clean datasets and create visualizations on those datasets. The locally saved files got loaded as DataFrames first and then those DataFrames got loaded as SqlLite tables named as 'us\_park\_species', 'park\_details' and 'species\_details'. These tables are getting further combined/merged using matching keys/data like Park Name, Species Name and final dataset combining all the 3 datasets is getting created based on matched values.  
In this project, the purpose of visualizations is to analyze and evaluate categories of species which are having biggest threat and if they are under the risk of endangered species. The 1st visualization (Bar Plot) is showing the count of species by biggest threat where 'Pesticide use' and 'Man/Human activity' are the topmost reasons. The 2nd visualization (Histogram Plot) is showing the distribution of Average Lifespan of species in years. It shows that average lifespan for majority of species is falling under 20-30 years of range. If those species are falling under biggest threat, then due to their low average lifespan there is a risk of possible extinction in near future. The 3rd visualization (Bar Plot) is showing which park has biggest threat and based on the generated graph, Sequioa and Kings Canyon National Park shows the highest. the 4th visualization (Pie Chart) is showing Proportion of Biggest Threat for each Category of Species, which interestingly shows that Birds are 72.3% victims of biggest threats (Pesticides, human activity).

The 5th visualization (Bar Plot) is generated to identify the number of recreation visitors for each park where we can assume that 'Sequioa and Kings Canyon National Park' shows highest threat, but number of recreation visitor is comparatively low with respect to other parks. Based on that, number of Recreation visitors doesn't seem to be impacting the cause of biggest threat for Sequioa. The 6th visualization (Count plot) is showing Species category distribution for each park where interesting fact to see is that 'Sequioa and Kings Canyon National Park' has majority 'Bird' category of species.  
 'Bird' category species is at highest threat based on previous plots and since Sequioa and Kings Canyon National Park is showing the highest count of species under Biggest Threat and got the 'Bird' as the most popular and highest count species, this park might be at a risk of endangered bird species.

**Limitations:**

**Inconsistent Species Information:**The final\_species.csv and final\_web\_species\_data.csv files have overlapping but not completely matching species-related data.

This inconsistency can lead to difficulties in accurately integrating and analyzing species information across the datasets.

**Lack of Comprehensive Threat Data:**The datasets provide some details about biggest\_threat and predators in the final\_species.csv file, but there is limited information on

other critical threats, such as habitat loss or climate change, which are essential for creating effective protection strategies.

**Data Completeness and Timeliness:**The final\_web\_species\_data.csv includes date\_established and recreation\_visitors information, which may not be frequently updated. Incomplete or outdated information can affect the accuracy of the AI/ML models in predicting current conservation status and population trends

Addressing these limitations can improve the effectiveness of the AI/ML project and

lead to better protection strategies for endangered species.

**Ethical Considerations:**

**Data Bias and Accuracy:** Inaccurate or biased data can lead to incorrect predictions, potentially diverting resources away from species in critical need or misinforming conservation strategies.

**Privacy Concerns:** Some datasets may include sensitive locations of endangered species. If this information is not handled correctly, it could lead to increased poaching or habitat destruction.

**Misuse of Predictions:** Predictions could be misinterpreted or misused by stakeholders, leading to harmful decisions or policies against the very species they aim to protect.

**Dependence on Technology:** Over-reliance on machine learning predictions could overshadow traditional conservation knowledge and practices, which are also crucial for species preservation.

**Resource Allocation:** Machine learning models might prioritize some species over others, influencing funding and conservation efforts and potentially neglecting species not deemed "important" by the model's criteria.

**Future Uses/Additional Applications:**

**Predictive Habitat Analysis**: By integrating species occurrence and habitat information,

the AI/ML models can predict potential habitats for endangered species. This can help

conservationists identify areas that need protection and restoration efforts.

**Visitor Impact Assessment:** Analyzing recreation visitor data alongside species data can provide insights into how human activities affect wildlife. This can lead to the development of strategies to minimize human impact on endangered species and their habitats.

**Conservation Prioritization:**Using AI/ML to assess the conservation status, abundance, and threats of various species, the project can prioritize conservation efforts for species most at risk. This can help allocate resources more effectively and ensure the most critical species receive the attention they need.

By leveraging these datasets, the AI/ML project can significantly enhance wildlife protection efforts and contribute to the preservation of endangered species.

**Recommendations:**

**Data Enrichment**: Enhance datasets by integrating additional external data sources,

such as climate data, habitat maps, and human activity data. This can provide a more

comprehensive view of the factors affecting endangered species and improve predictive accuracy.

**RealTime Monitoring**: Implement realtime data collection and monitoring systems to keep the datasets updated. This can involve using IoT sensors, satellite imagery,

and citizen science contributions to ensure the AI/ML models have the most current information for accurate analysis and decision-making.

**Interdisciplinary Collaboration**: Collaborate with ecologists, conservationists, and data

scientists to refine AI/ML models and ensure they are grounded in ecological principles. This can help in developing targeted conservation strategies that are both scientifically

 sound and practically feasible.

By adopting these recommendations, the project can enhance its effectiveness and

contribute significantly to the protection of endangered species.

**Ethical Assessment:**

Ensuring fairness in model predictions across different species, ecosystems, and demographics is crucial and efforts should be made to mitigate bias and promote equitable outcomes. During data transformation, some of the NaN or empty values got converted into 'Unknown' value for consistency or has been removed and filtered out . But there is a possibility that some crucial species information might have lost during this clean up as they did not have sufficient information recorded. If the data used for training is biased (e.g., underrepresents certain species or habitats), the model may produce biased predictions.

Based on visualizations, one of the observation came up that 'Man/Human Activity' is one of the topmost reasons of biggest threat for species. In that case, number of recreation visitors also play an important role because a species habitat can be impacted with increasing visitors so the data need to be accurate and should get updated with recent information for this project. Accurate information or missing data can lead to inappropriate results.

Another important ethical implication is Transparency on how the data collection process is taking place, including what data is being collected, how it will be used, and for what purpose considering all the guidelines for website and API calls. This data should not be used inadvertently which can cause intentional impact to tourism or any decision regarding national parks or any species.

**Challenges/Issues:**

When working with data for endangered species in the context of machine learning several concerns and challenges may arise:

**Data Availability**: API data on endangered species can be scarce or incomplete due to the limited number of individuals and the difficulty in tracking and monitoring these species in their natural habitats. After running the API for a few animals, not every output provides all the information required in the final dataset.

**Temporal and Spatial Variability**: Endangered species data can be highly variable over time and space, requiring complex models to understand patterns of behavior, migration, and population dynamics.

**Impact of Climate Change**: The effects of climate change on species habitats, migration patterns, and population dynamics add another layer of complexity to modeling efforts.

**Integration with Conservation Strategies**: The ultimate goal is to aid conservation efforts, which requires models not only to predict but also to prescribe actionable strategies. This requires a deep understanding of the ecological, social, and economic

**References:**

Here are some additional sources that can be used to validate results and support the Wildlife Protection of Endangered Species project using AI/ML:

1. **Scientific Journals and Publications**

Sources: Journals such as "Conservation Biology," "Biodiversity and Conservation," and "Ecological Applications."

Description: These journals publish peer-reviewed research articles on wildlife conservation, habitat management, and the application of AI/ML in ecological studies. They provide credible scientific evidence and methodologies to support the project.

1. **International Union for Conservation of Nature (IUCN) Red List**

Source: IUCN Red List of Threatened Species

Description: The IUCN Red List is a comprehensive resource that provides information on the conservation status of species worldwide. It includes data on species populations, threats, and conservation measures, which can be used to validate species-specific results.

Website: IUCN Red List

1. **World Wildlife Fund (WWF) Reports**

Source: WWF

Description: WWF publishes reports and white papers on various aspects of wildlife conservation, including endangered species protection, habitat preservation, and the impact of climate change. These reports offer valuable insights and real-world case studies.

Website: World Wildlife Fund

1. **United Nations Environment Programme (UNEP)**

Source: UNEP

Description: UNEP provides reports, data, and policy recommendations on global environmental issues, including biodiversity and endangered species conservation. Their publications can help validate the project's methodologies and findings.

Website: UNEP

1. **Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES)**

Source: CITES

Description: CITES maintains a database of species that are regulated under international trade agreements. It provides information on species protection measures and trade regulations, which can be used to support conservation efforts.

Website: CITES

1. **National Geographic**

Source: National Geographic

Description: National Geographic publishes articles, documentaries, and research on wildlife conservation and endangered species. Their content often includes expert insights and field research that can complement the project's data and analysis.

Website: National Geographic

These sources offer credible and comprehensive information that can validate the results and support the Wildlife Protection of Endangered Species project using AI/ML.

10 user questions regarding the Wildlife Protection of Endangered Species AI/ML project?

1. Whats the benefit of Wildlife Protection of Endangered  to general public?

Ans: Protecting endangered wildlife helps maintain biodiversity, ensuring balanced ecosystems. It promotes economic growth through ecotourism, creating jobs and supporting local communities. Additionally, healthy ecosystems provide vital services like clean air, water, and disease regulation, benefiting public health.

1. What changes were made to the data?

Ans: No major changes has been done to data in this milestone project. All the 3 datasets got loaded into SqlLite tables and during joining some redundant columns got filtered out in the final dataset. For consistency, some NaN/None values got converted into 'Unknown' across the final dataset.

1. How do you handle missing or inconsistent data across the different datasets?

Ans: To handle missing or inconsistent data, first, identify and fill in missing values using techniques like mean imputation or interpolation, or flag them for review. Next, standardize data formats, units, and naming conventions across datasets. Finally, cross-verify data to resolve any discrepancies, ensuring consistent and accurate information. These steps help maintain data integrity.

1. Are there any legal or regulatory guidelines for your data or project topic?

Ans: Not required as the data is collected from a valid API source using secret API key and includes species information only.

1. How do you measure the effectiveness and accuracy of the AI/ML models in protecting endangered species?

Ans: To measure the effectiveness and accuracy of AI/ML models in protecting endangered species, monitor key metrics such as prediction accuracy, precision, and recall. Additionally, track real-world outcomes like the reduction in poaching incidents and improvements in species population trends. Regularly validate model predictions against field data and adjust models as necessary for continuous improvement.

1. What risks could be created based on the transformations done?

Ans: No major transformations have been done in this milestone project but for visualizations one big chunk of data has been filtered out for biggest threat as 'Unknown'. Since the visualizations got generated on limited data, there is a possibility of biased predictions or incomplete analysis.

1. Did you make any assumptions in cleaning/transforming the data?

Ans: Yes, for visualization a big chunk of data got filtered out with 'Unknown' value as biggest threat, so all the assumptions were made on limited dataset. The limited dataset (new\_df) is containing only 191 rows of species data.

1. Was your data acquired in an ethical way?

Ans: The data has been acquired in ethical ways as all sources are trusted and park information is managed by National Park Services in United States. However, the time and effort spent on species inventories varies from park to park, which may result in data gaps. Species taxonomy changes over time and reflects regional variations or preferences; therefore, records may be listed under a different species name.

1. How would you mitigate any of the ethical implications you have identified?

Ans: Since the final dataset has limited amount of data which can be used for training purposes of the model but further data collection from trusted sources can provide better results and more insights. In further analysis, if species data gets outdated then a mechanism need to be implemented where real time data on certain park information will get feeded from other credible sources.

1. How can stakeholders and conservationists’ access and use the insights generated by the AI/ML models?

Ans: Stakeholders and conservationists can access insights from AI/ML models via user-friendly dashboards, reports, and data visualization tools. These platforms provide actionable recommendations, real-time alerts, and comprehensive analytics. By utilizing these insights, they can make informed decisions, prioritize conservation efforts, and monitor the effectiveness of their strategies.