**Group 370 : INaturalist Observation Data Prediction Model**

Group 370 : Making a prediction model using past observation data to determine the month a species is most likely observed during.

Your submissions:

* 370\_Report.pdf
* 370\_Codes.ipynb (with necessary comments)

Notes

* No extension to the deadline
* Each team can only submit one copy by a single member, just list all of your members in the report
* use RED font for the parts that you revised according to the feedbacks in your presentation

|  |  |  |
| --- | --- | --- |
| First Name | Last Name | Email address |
| Piotr | Rajchel | prajchel@hawk.iit.edu |

Table of Contents

[**1. Introduction** 2](#_Toc56029461)

[**2. Data** 2](#_Toc56029462)

[**3. Problems and Solutions** 2](#_Toc56029463)

[**4. KDD** 2](#_Toc56029464)

[4.1. Data Processing 2](#_Toc56029465)

[4.2. Data Mining Methods and Processes 2](#_Toc56029466)

[**5. Evaluations and Results** 2](#_Toc56029467)

[5.1. Evaluation Methods 2](#_Toc56029468)

[5.2. Results and Findings 2](#_Toc56029469)

[**6. Conclusions and Future Work** 2](#_Toc56029470)

[6.1. Conclusions 2](#_Toc56029471)

[6.2. Limitations 2](#_Toc56029472)

[6.3. Potential Improvements or Future Work 2](#_Toc56029473)

# **1. Introduction**

Last semester I made a basic Neural Network model to identify different bird species based on images. This semester, I’m keeping the environmental theme and using INaturalist Species Observation data. After thinking about applications and uses for this data in machine learning, I realized AI can’t randomly generate meaningful observation data, so instead I’ll be using it to determine what month a given species is most likely to be found during. I’ll enter a random species that resides in the states of Illinois, Michigan, Wisconsin, and Indiana, which means it should already exist in the training set, and it will determine what month was most likely for me to find it during. In other words, this information could be used to help you know what month the best month is to go searching for this species.

Another possible idea was to determine an average location or range, but locational data predicted by an ai isn’t that useful in a scenario like this.

# **2. Data**

My dataset was gathered through INaturalist’s Export to CSV function in their observation search. The database is massive, so I decided to narrow down the data to between 2015-2022, to only include amphibians and reptiles, and only in the states of Illinois, Michigan, Wisconsin, and Indiana.

When it comes to preprocessing the data, my first step before even downloading the data was deciding which columns I needed. The database provides a huge number of rows and columns, but many are for storage purposes, so rather than removing them after downloading the database, I did it ahead of time on the website. I also did not include data below “research grade”, meaning that each observation has been reviewed by peers. These are the columns I chose to leave:

A screenshot of a computer

Description automatically generated with medium confidence

# In the code, additional pre-processing included dropping the columns of id, URL, and image URL, as they were reference columns that were unnecessary in creating the data model. Additional dropped or modified rows included:

* **Captive\_Cultivated:** This row is only for when species are recorded that are raised by people, and not wild observations. Not only did I not want that data, but there was also not a single instance of a TRUE row in all the 71k+ rows, so the column was dropped.
* **Positional\_Accuracy:** Geospatial data isn’t too important when I’m looking specifically at observation month, and even if I was using it more, I’m keeping latitude and longitude, and I don’t need exact map location, so this column was removed.
* **Observed\_on\_String:** This column was the same as observed\_on, except with extra time info added onto the dates, so I removed it to reduce redundancy.
* **Time\_Observed\_at:** This column had more info than was useful, and to reduce redundancy, I used regex to remove anything but the 00:00:00 format timestamps and used fillna method ffill to fill missing values.
* **Time\_Zone:** Time values aren’t crucially important, so I used fillna method ffill as well to fill missing values.
* **Place\_Guess:** Replaced missing values with “None”, as there was no guess for the location.
* **Species\_Guess:** Replaced missing values with “None”, as there was no guess for the species.

# **3. Problems and Solutions**

Describe the problems to be solved

Note: they are problems, not the solutions. For example, hypothesis testing, ANOVA, classification, clutering, etc, are all solutions

# **4. KDD**

## 4.1. Data Processing

xxxxx

## 4.2. Data Mining Methods and Processes

xxxx

# **5. Evaluations and Results**

## 5.1. Evaluation Methods

xxxx

## 5.2. Results and Findings

xxx

# **6. Conclusions and Future Work**

## 6.1. Conclusions

xxx

## 6.2. Limitations

xxx

## 6.3. Potential Improvements or Future Work

Xxx