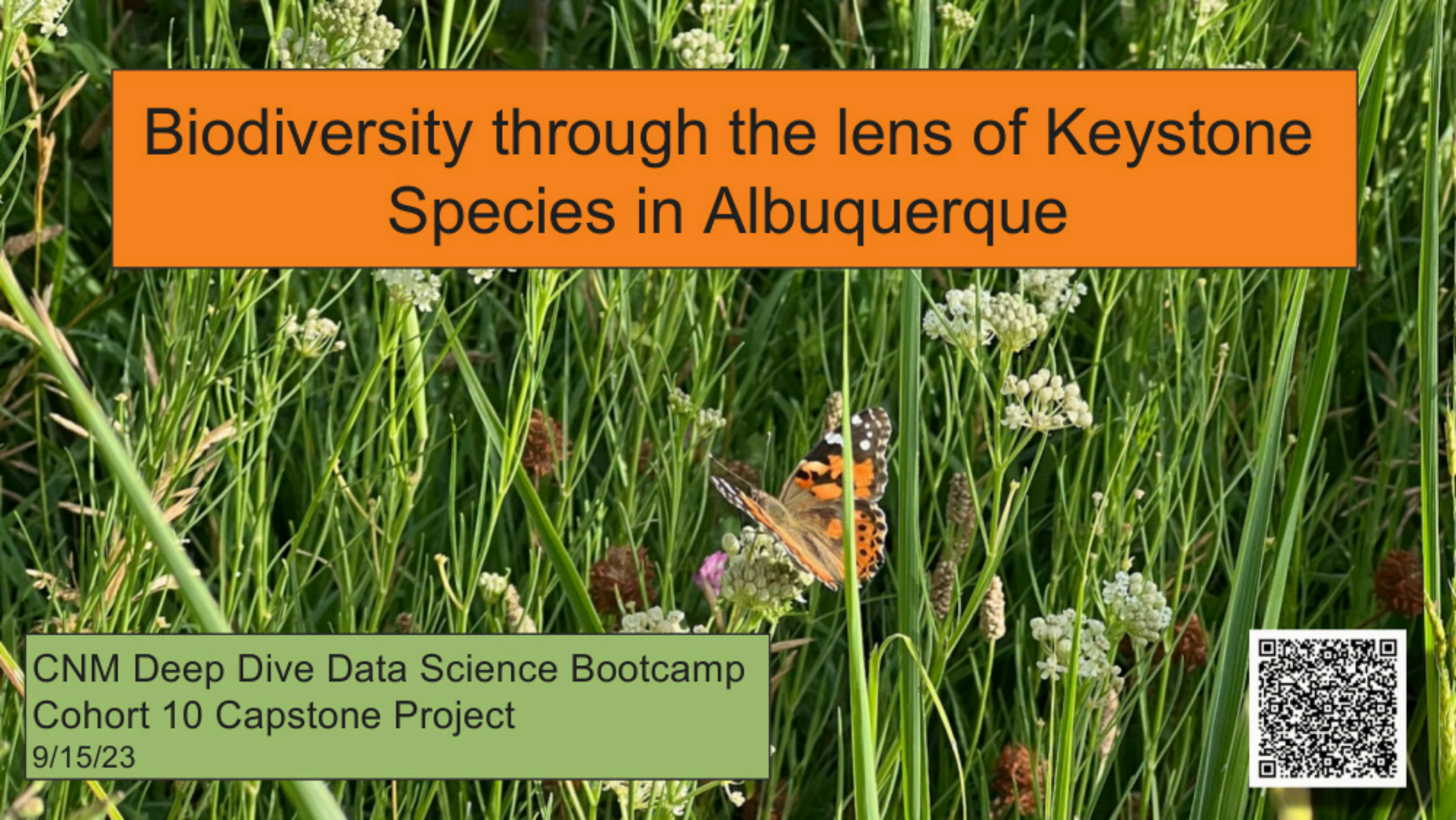
Biodiversity through the lens of Keystone Species in Albuquerque



CNM Deep Dive Data Science Bootcamp
Cohort 10 Capstone Project
9/15/23

Problem Definition

Citizen science may be an under recognised data source. With tree-based and linear regression modeling we are hoping to show how citizen science can be used to understand the relationship between soil type, land cover classification and the predictive factors for keystone species in the urban environment of Albuquerque, New Mexico.

Concept of Keystone Species

- Robert Paine experiment
  - Early 1960's
  - Conducted at Mukkaw Bay, WA
  - Similar observations made by other Ecologists
- Organisms that keep ecosystem in equilibrium

Ecoregions of New Mexico

- Arizona / New Mexico Plateau
- Chihuahuan Desert
- Southern Rockies
- Great Plains
- Arizona / New Mexico Mountains
- Madrean Archipelago

Keystone Species - Arizona / NM Plateau

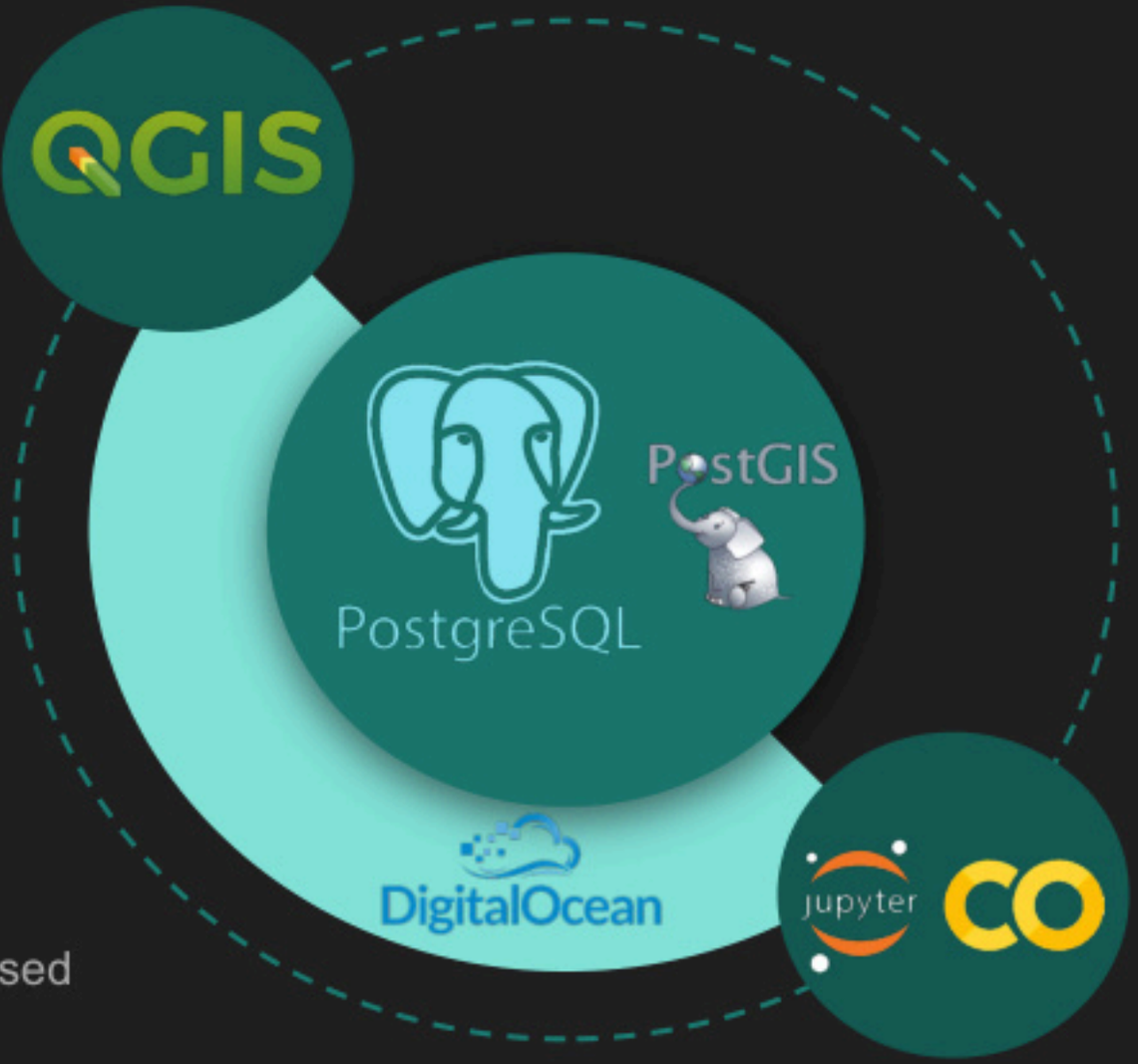
- Predator
  - Coyote
- Prey
  - Cottontail Rabbit
- Ecosystem Engineer
  - Beaver
  - Prairie Dog
- Mutualist
  - Hummingbird
  - Bee
- Plants
  - Coyote Willow
  - Chamisa
  - Cottonwood

Data Sources and Infrastructure

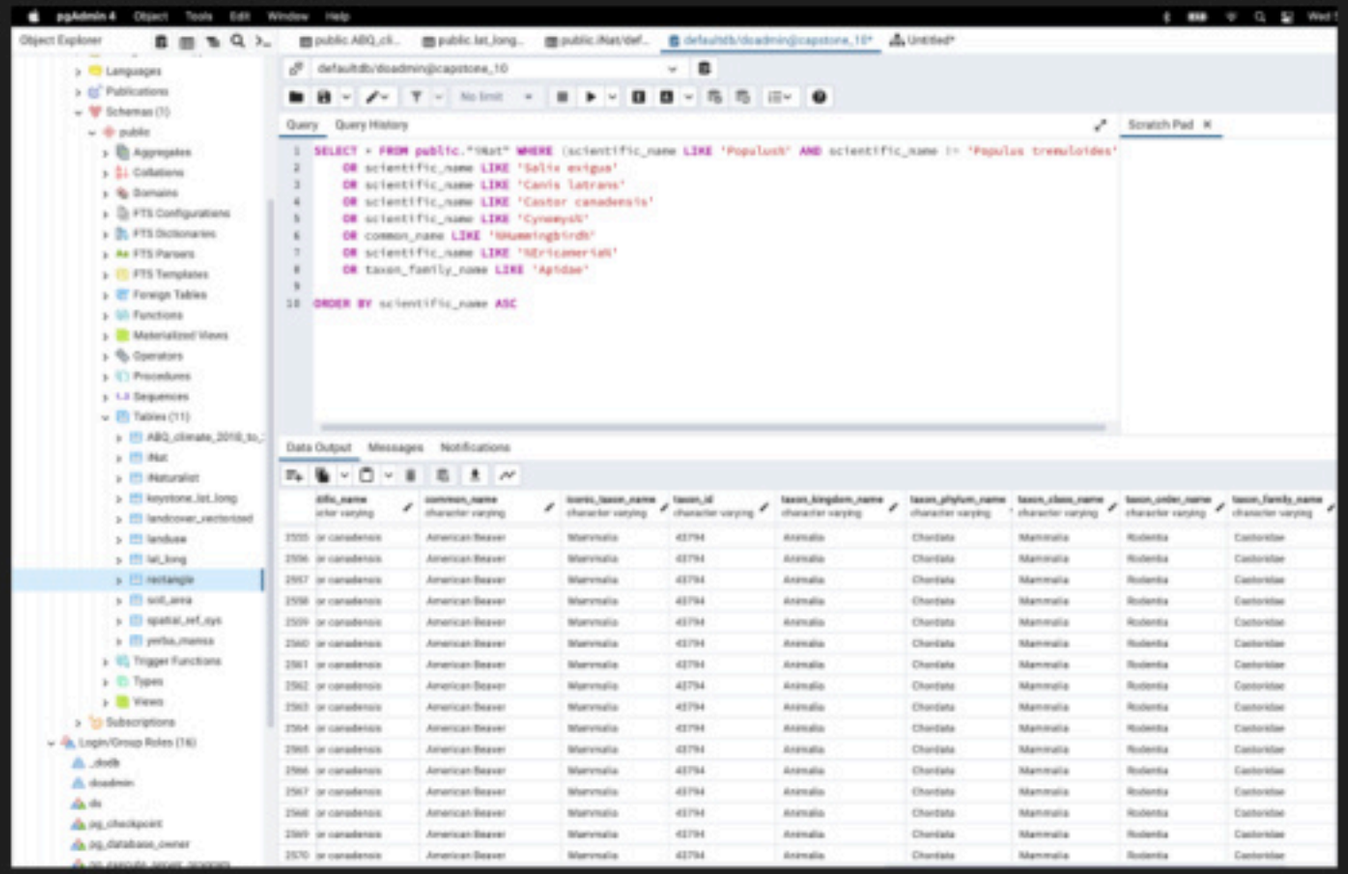
iNaturalist: an online social network of people sharing biodiversity information to help each other learn about nature.

USGS Multi-Resolution Land Characteristics Consortium (MRLC) land cover data from latest 2021 survey

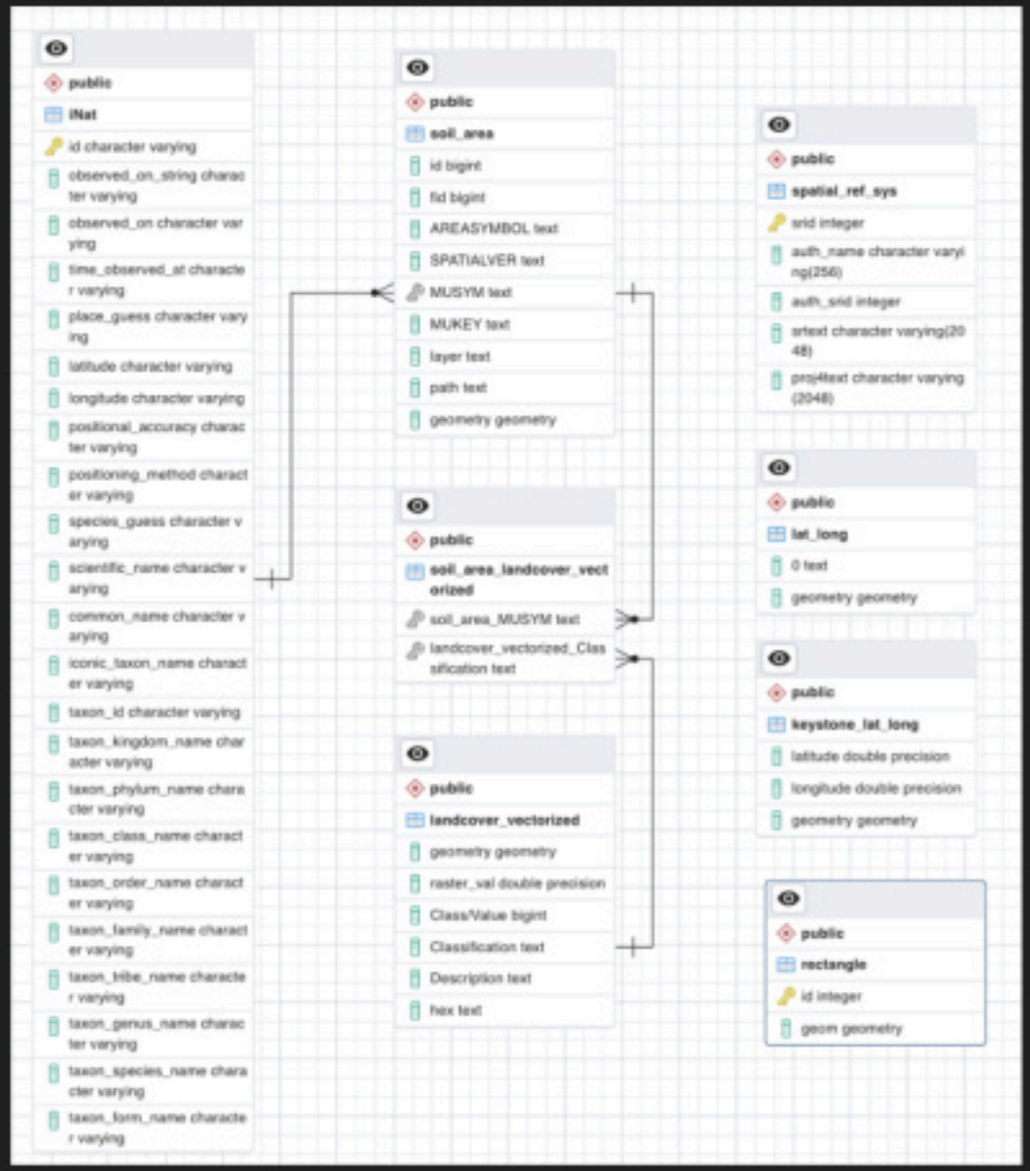
USDA Web Soil Survey: tiled area of interest over City of Albuquerque processed in QGIS



pgAdmin 4



Query Tool with SQL Console



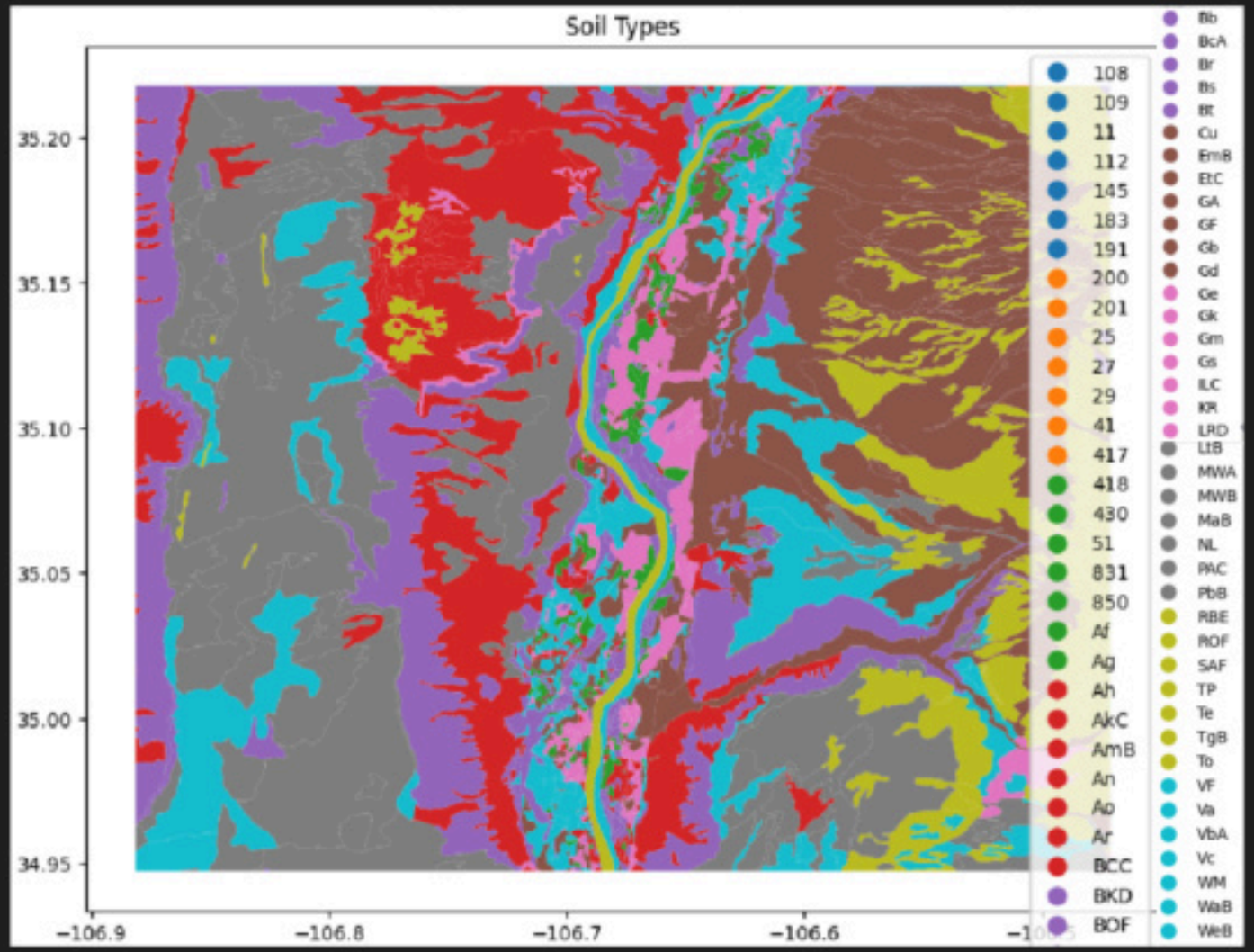
Entity Relationship Diagram (ERD)

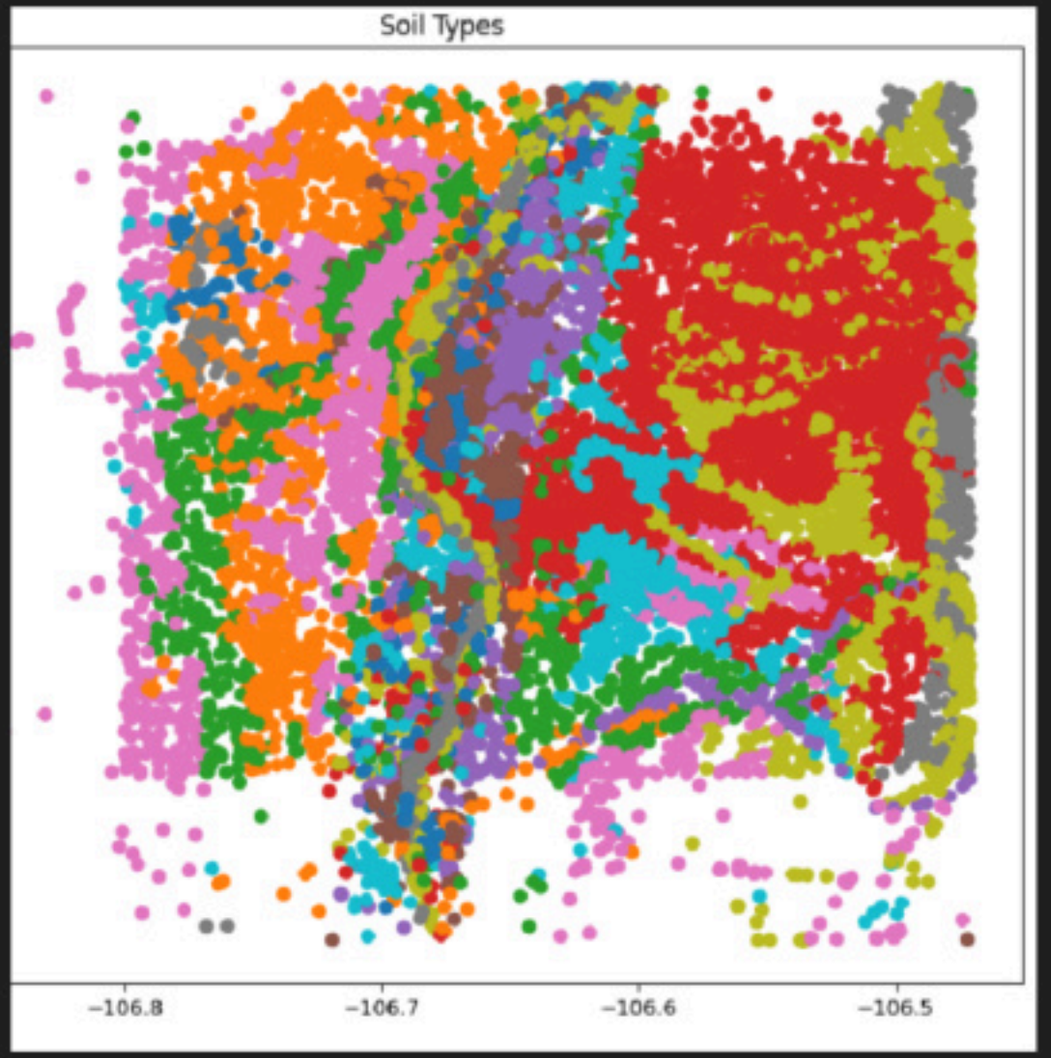
Technology Overview

Machine Learning Techniques Used

- Decision Trees
- Random Forest
- Logistic Regression
- K-Means Clustering
- XGBoost
  - Feature Engineering
- Hierarchical Clustering
- RMSE

Feature Engineering: Soil Type

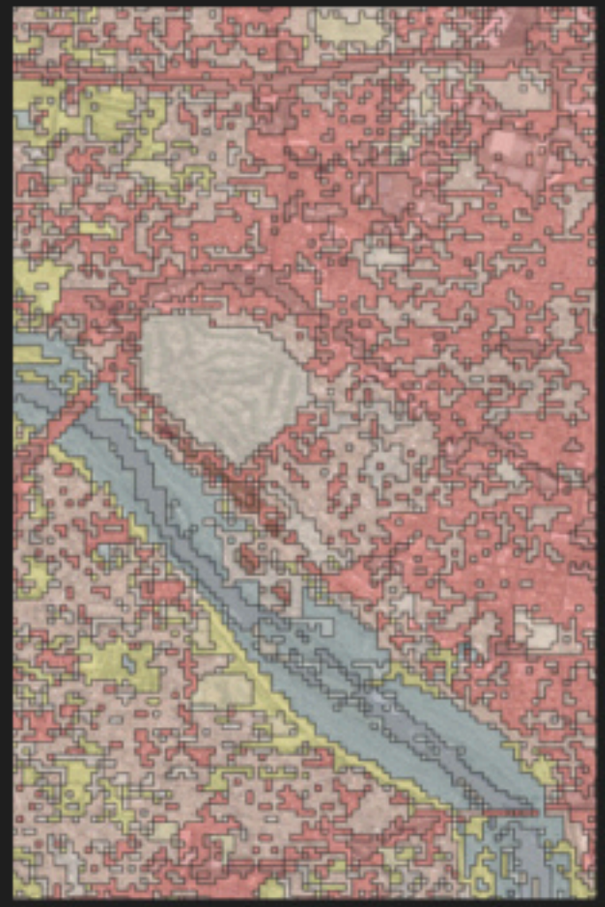




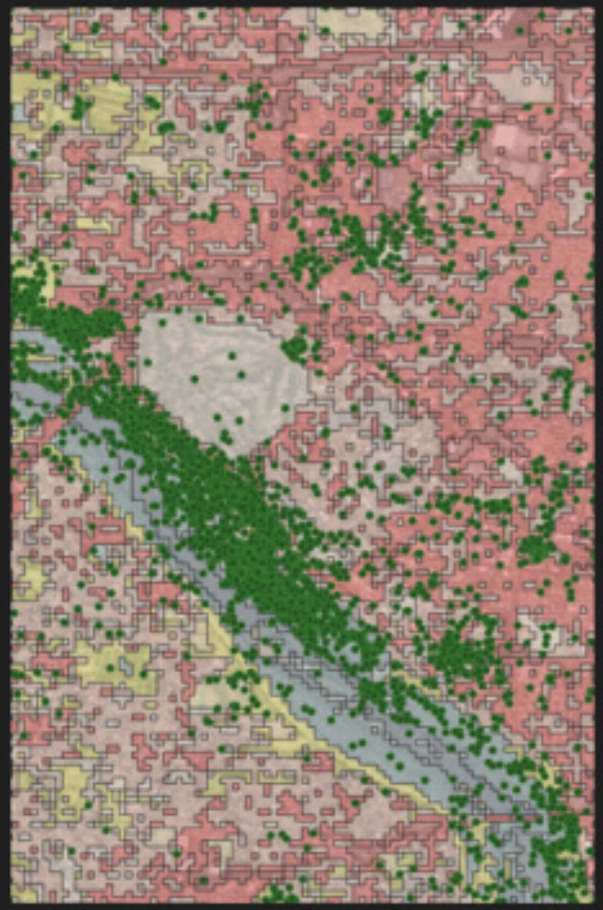
Feature Engineering: Land Cover

Aerial Image (Raster)

Land Cover (Raster)

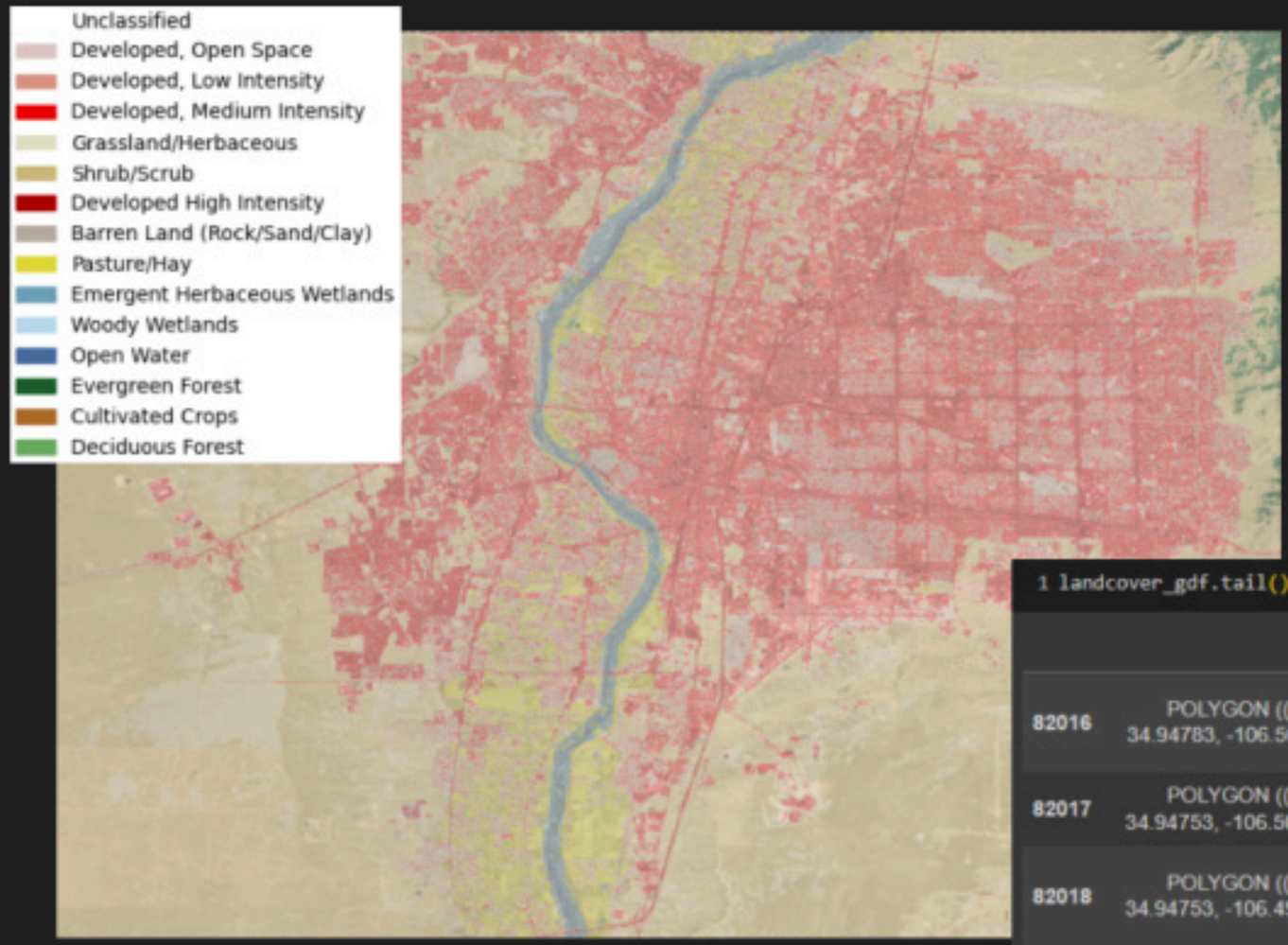


Land Cover (Vectorized)



iNaturalist Observations

Feature Engineering: Land Cover



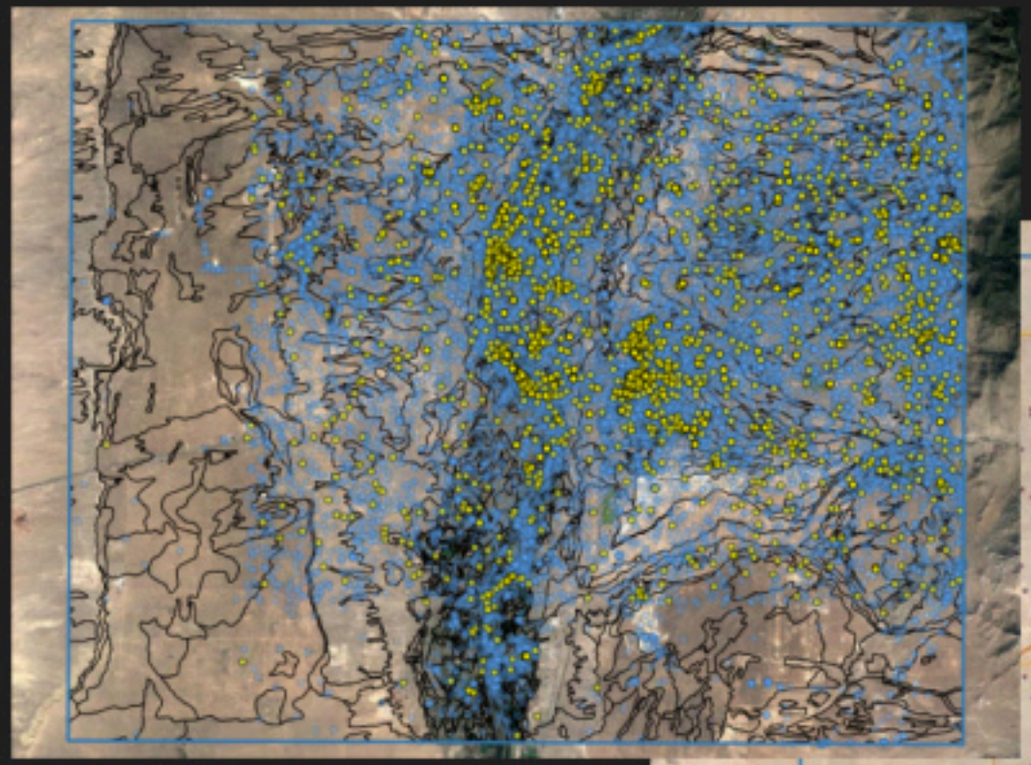
City of Albuquerque 2021 Land Cover

Land Cover Classification to GeoDataFrame

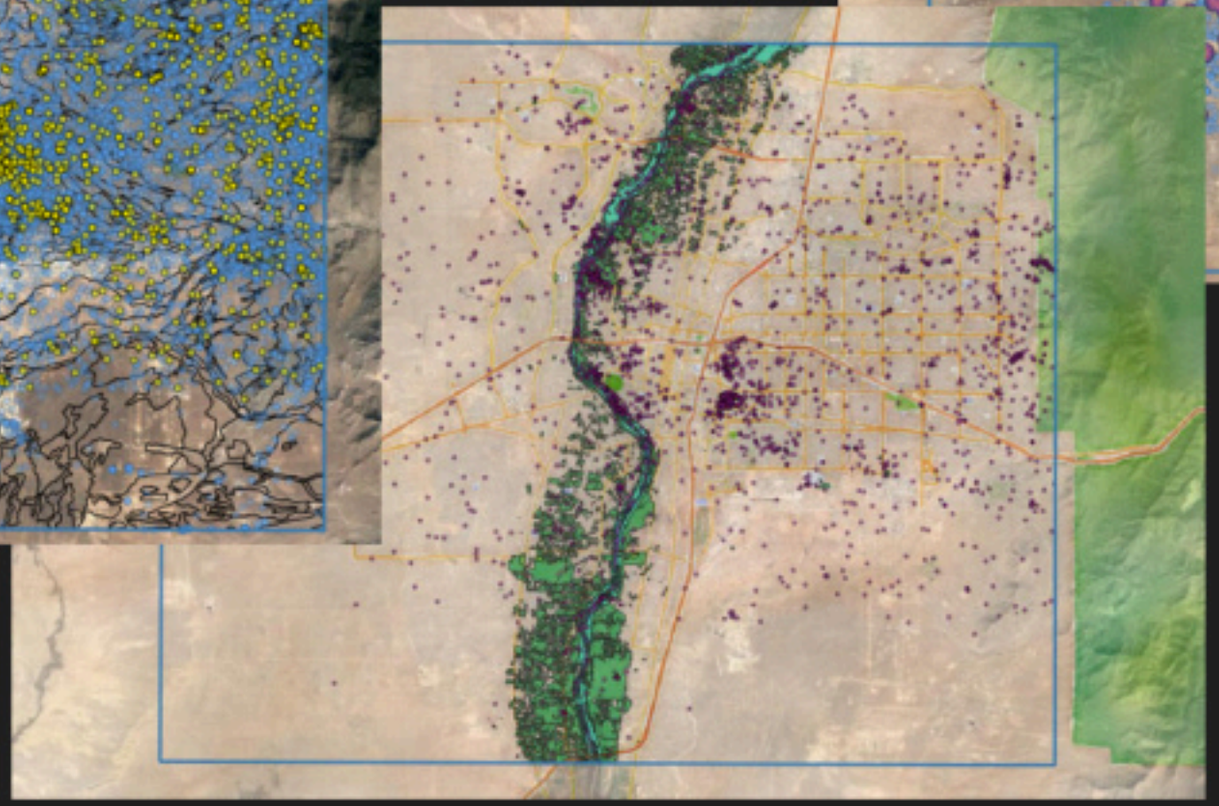
```
1 landcover_gdf.tail()
```

	geometry	raster_val	Class/Value	Classification	Description	hex
82016	POLYGON ((-106.50144 34.94783, -106.50114 34.9...	71.0	71	Grassland/Herbaceous	areas dominated by graminoid or herbaceous veg...	#dfdfc2
82017	POLYGON ((-106.50024 34.94753, -106.50024 34.9...	52.0	52	Shrub/Scrub	areas dominated by shrubs; less than 5 meters ...	#ccb879
82018	POLYGON ((-106.49934 34.94753, -106.49934 34.9...	71.0	71	Grassland/Herbaceous	areas dominated by graminoid or herbaceous veg...	#dfdfc2
82019	POLYGON ((-106.48465 34.95143, -106.48375 34.9...	71.0	71	Grassland/Herbaceous	areas dominated by graminoid or herbaceous veg...	#dfdfc2
82020	POLYGON ((-106.48855 35.06504, -106.48795 35.0...	52.0	52	Shrub/Scrub	areas dominated by shrubs; less than 5 meters ...	#ccb879

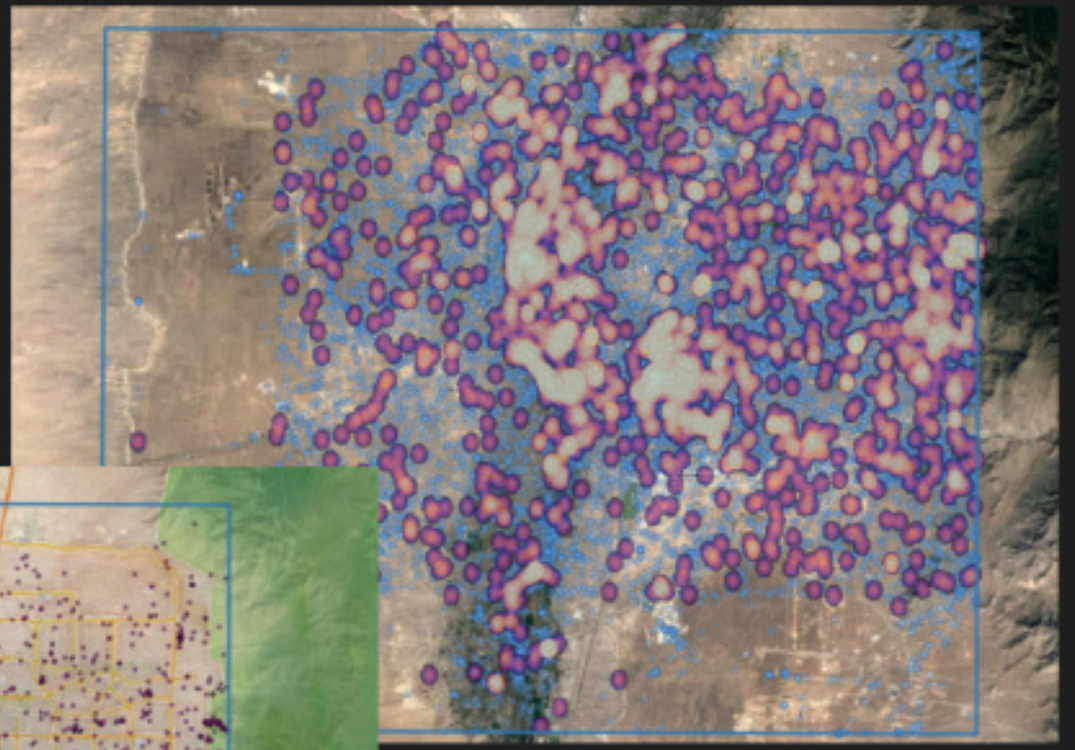
Geospatial EDA



Soil Polygons with iNaturalist Observations



Keystone Species vs. Riparian Land Cover

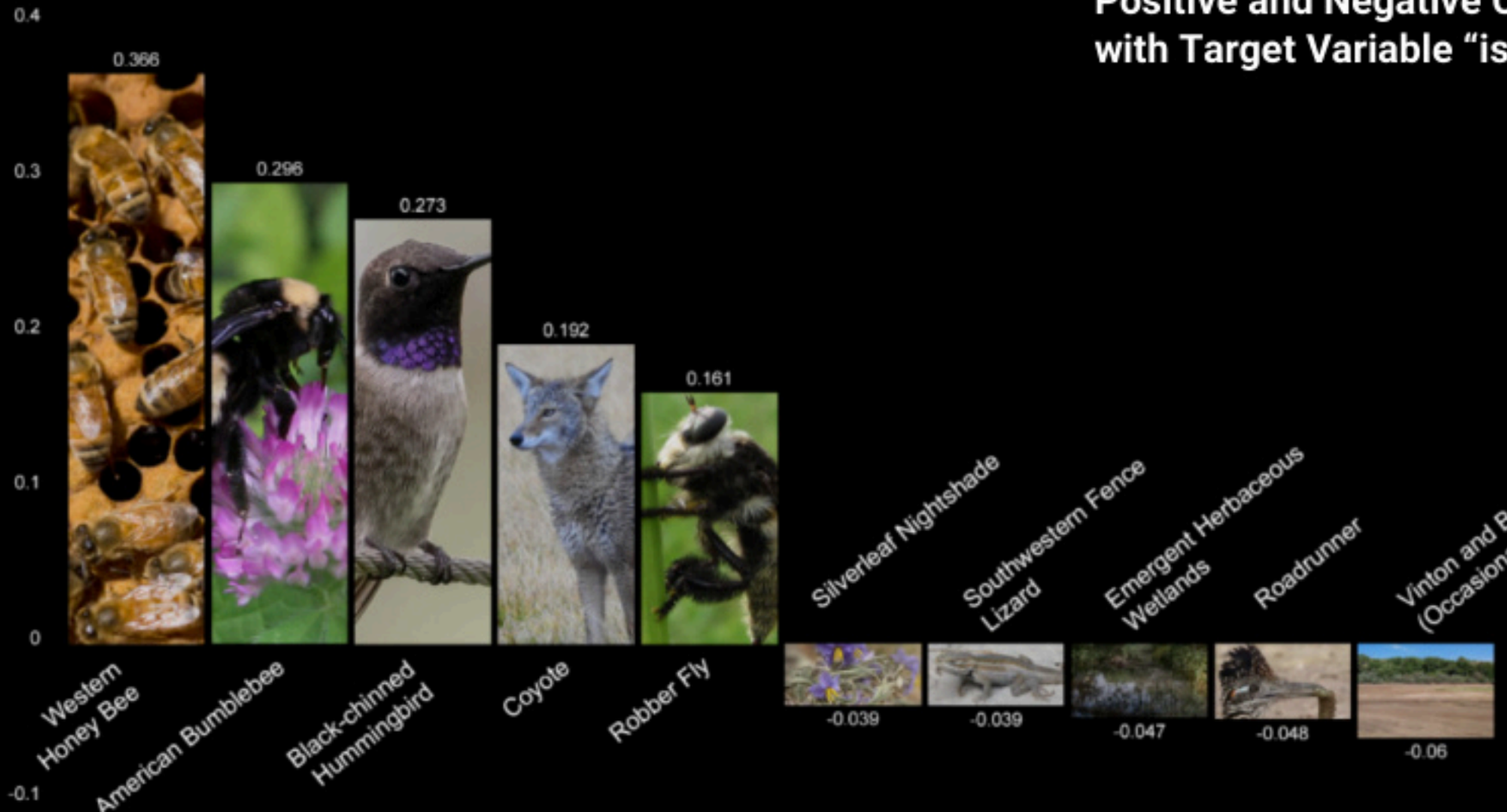


Heat Map Showing Distribution of Keystone Species

QGIS

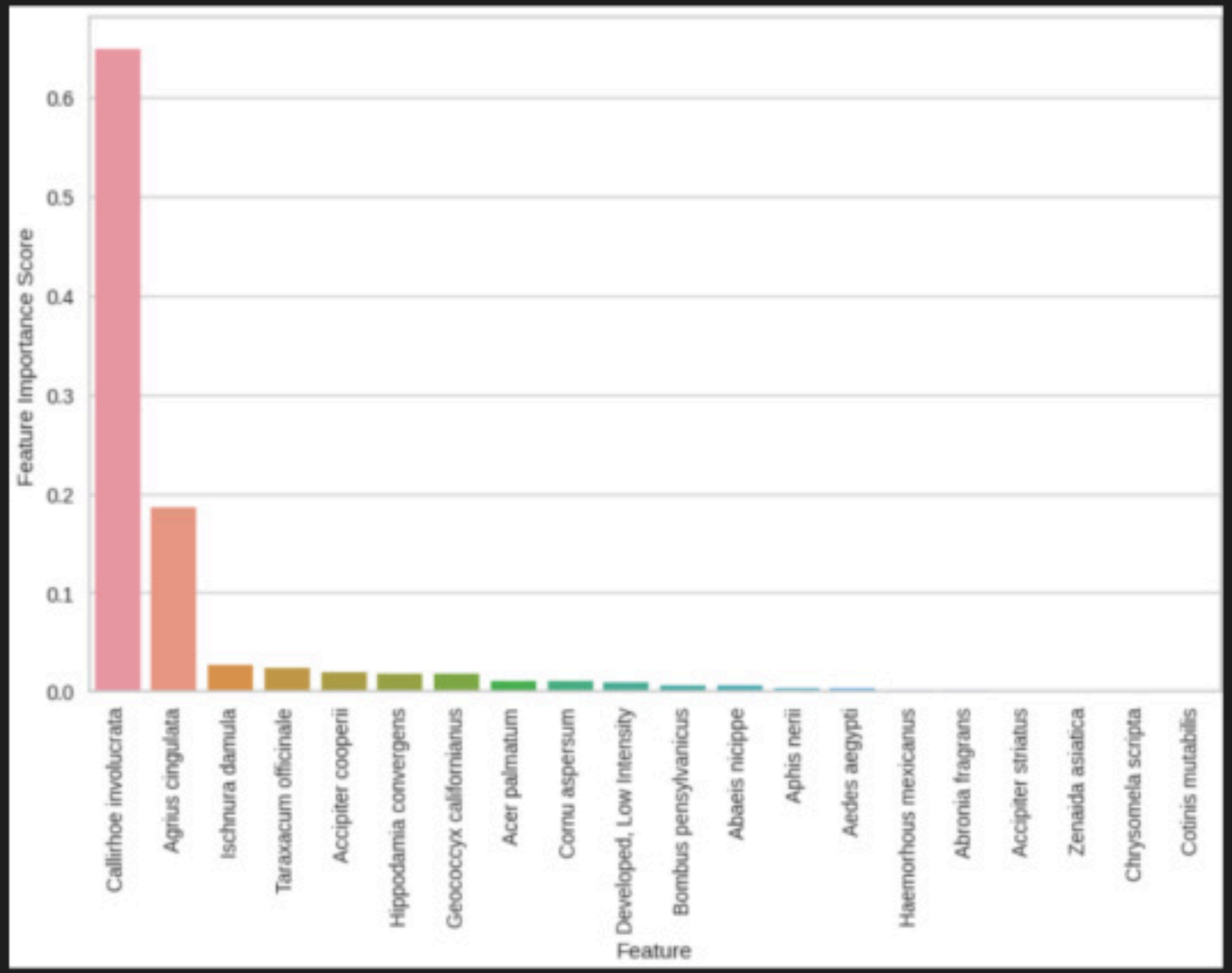
Combined Features Analysis (Excluding Decision Tree)

Positive and Negative Correlation with Target Variable “is_keystone”



This updated bar chart combines the top 5 positively and negatively correlated features with the target variable 'is_keystone', excluding the decision tree important features. The x-axis labels are directly below the bars for easier identification.

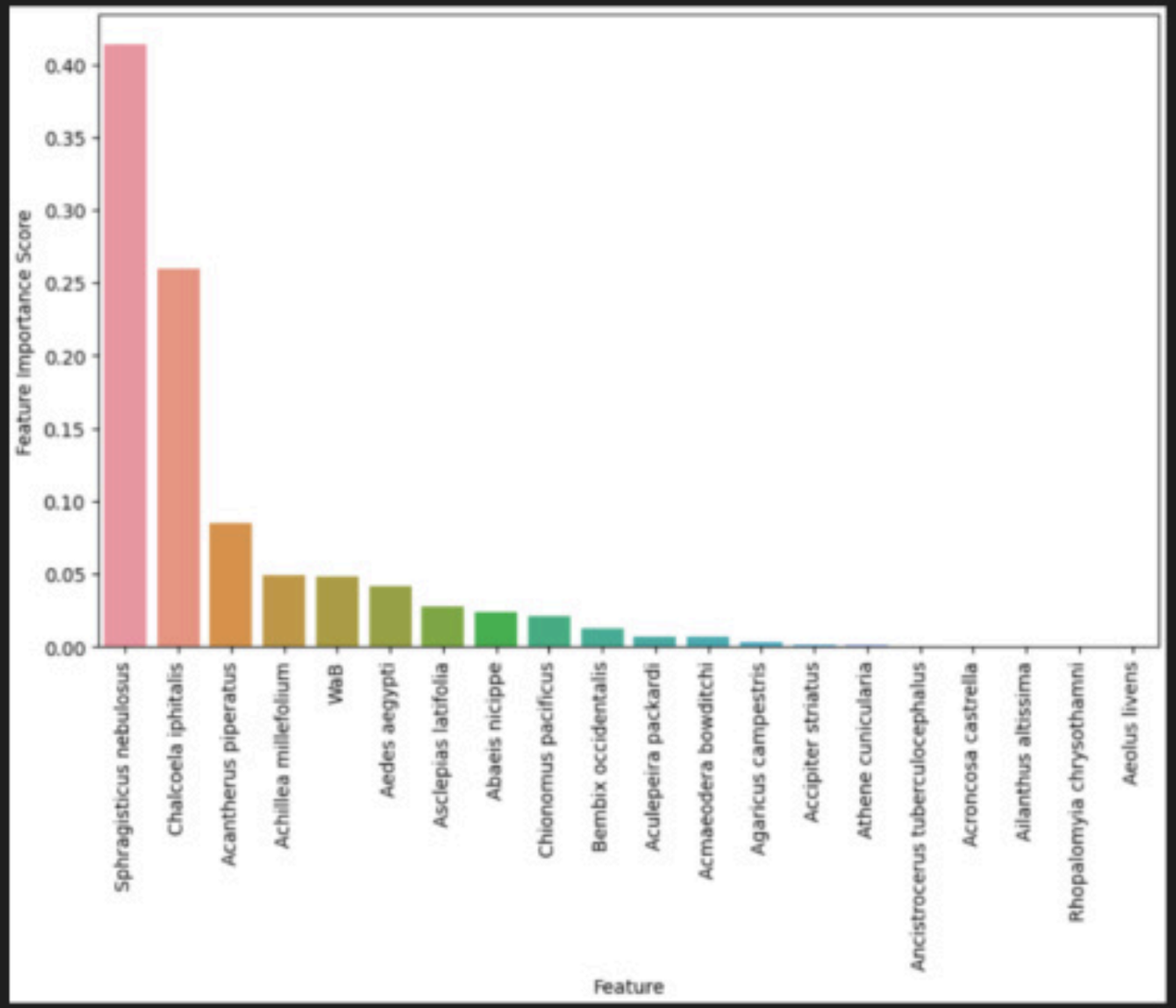
Honey Bees! RMSE: 0.09681156163532739



- The most important feature for the presence of honey bees in XGBoost is unsurprisingly a flower

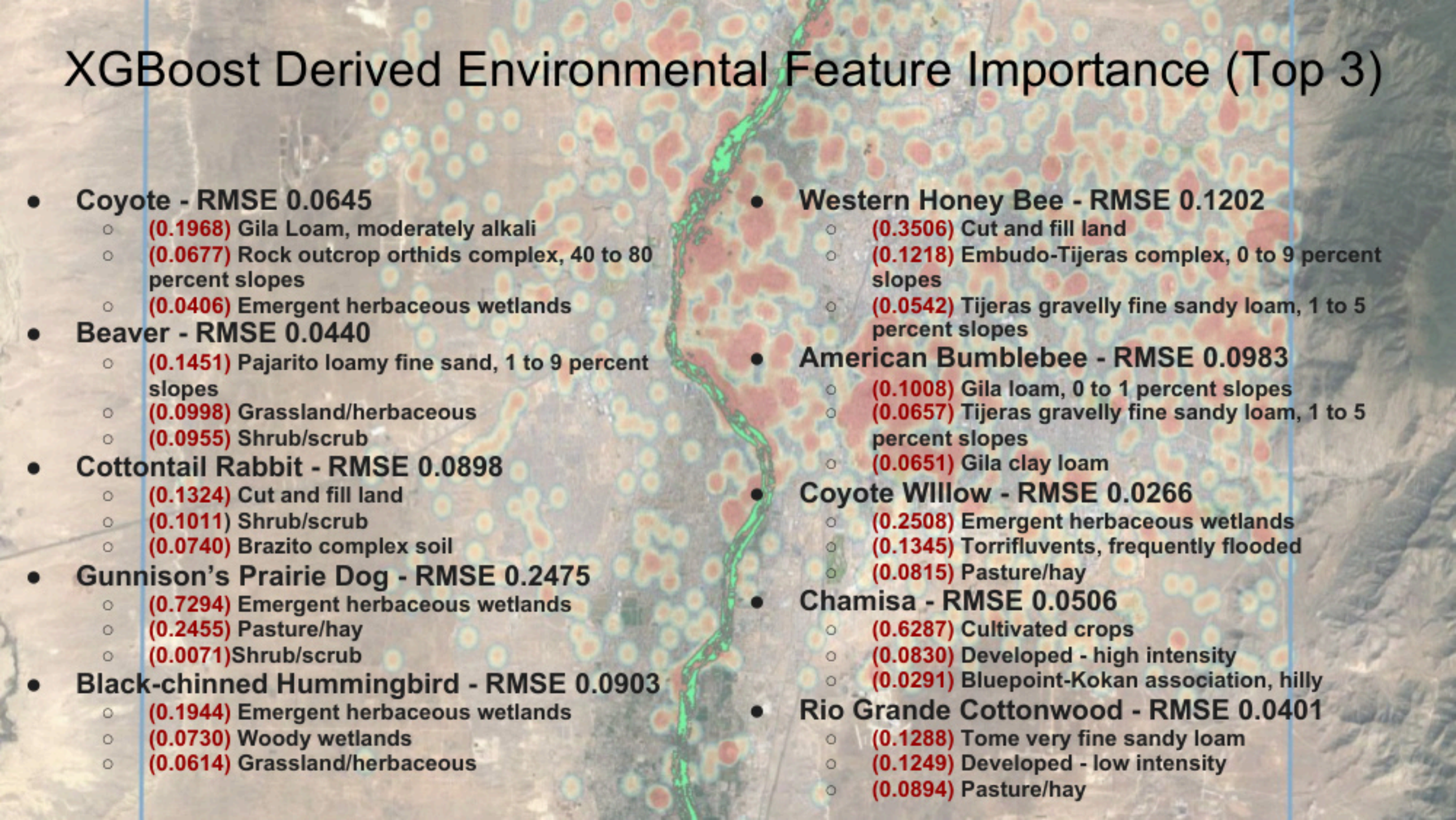
Purple Poppy-Mallow, courtesy of https://en.wikipedia.org

Prairie Dogs! RMSE: 0.1611310832860667



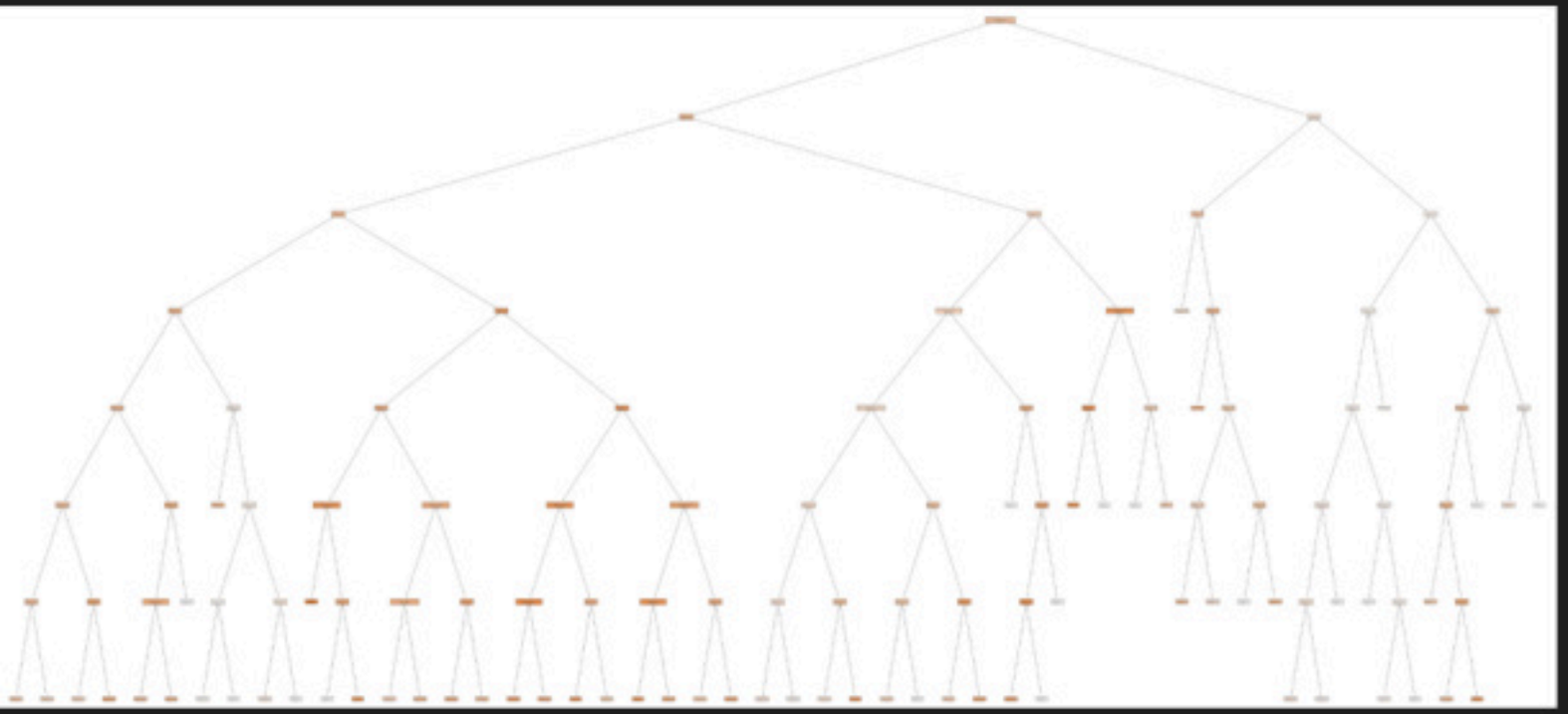
Seed bug courtesy of https://www.britishbugs.org.uk

XGBoost Derived Environmental Feature Importance (Top 3)



- **Coyote - RMSE 0.0645**
  - (0.1968) Gila Loam, moderately alkali
  - (0.0677) Rock outcrop orthids complex, 40 to 80 percent slopes
  - (0.0406) Emergent herbaceous wetlands
- **Beaver - RMSE 0.0440**
  - (0.1451) Pajarito loamy fine sand, 1 to 9 percent slopes
  - (0.0998) Grassland/herbaceous
  - (0.0955) Shrub/scrub
- **Cottontail Rabbit - RMSE 0.0898**
  - (0.1324) Cut and fill land
  - (0.1011) Shrub/scrub
  - (0.0740) Brazito complex soil
- **Gunnison's Prairie Dog - RMSE 0.2475**
  - (0.7294) Emergent herbaceous wetlands
  - (0.2455) Pasture/hay
  - (0.0071)Shrub/scrub
- **Black-chinned Hummingbird - RMSE 0.0903**
  - (0.1944) Emergent herbaceous wetlands
  - (0.0730) Woody wetlands
  - (0.0614) Grassland/herbaceous
- **Western Honey Bee - RMSE 0.1202**
  - (0.3506) Cut and fill land
  - (0.1218) Embudo-Tijeras complex, 0 to 9 percent slopes
  - (0.0542) Tijeras gravelly fine sandy loam, 1 to 5 percent slopes
- **American Bumblebee - RMSE 0.0983**
  - (0.1008) Gila loam, 0 to 1 percent slopes
  - (0.0657) Tijeras gravelly fine sandy loam, 1 to 5 percent slopes
  - (0.0651) Gila clay loam
- **Coyote WIllow - RMSE 0.0266**
  - (0.2508) Emergent herbaceous wetlands
  - (0.1345) Torrifluvents, frequently flooded
  - (0.0815) Pasture/hay
- **Chamisa - RMSE 0.0506**
  - (0.6287) Cultivated crops
  - (0.0830) Developed - high intensity
  - (0.0291) Bluepoint-Kokan association, hilly
- **Rio Grande Cottonwood - RMSE 0.0401**
  - (0.1288) Tome very fine sandy loam
  - (0.1249) Developed - low intensity
  - (0.0894) Pasture/hay

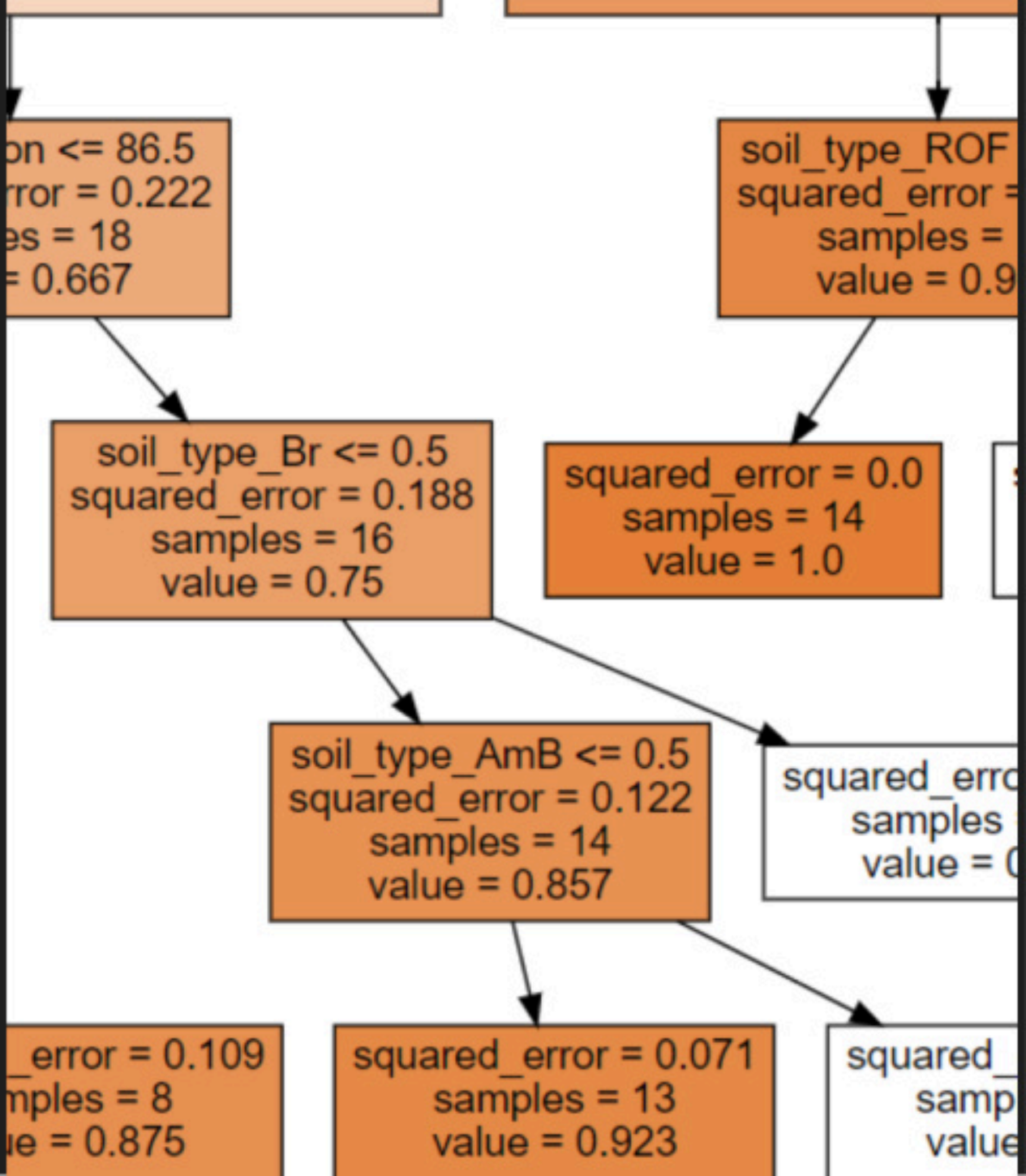
Regression Comparisons



- iNaturalist - Soil - Land Use
  - Filtered for target keystone genera - target
- XGBoost
  - Calculated RMSE and top 20 most important features
- Decision Tree
  - Filtered out all features except for top 20

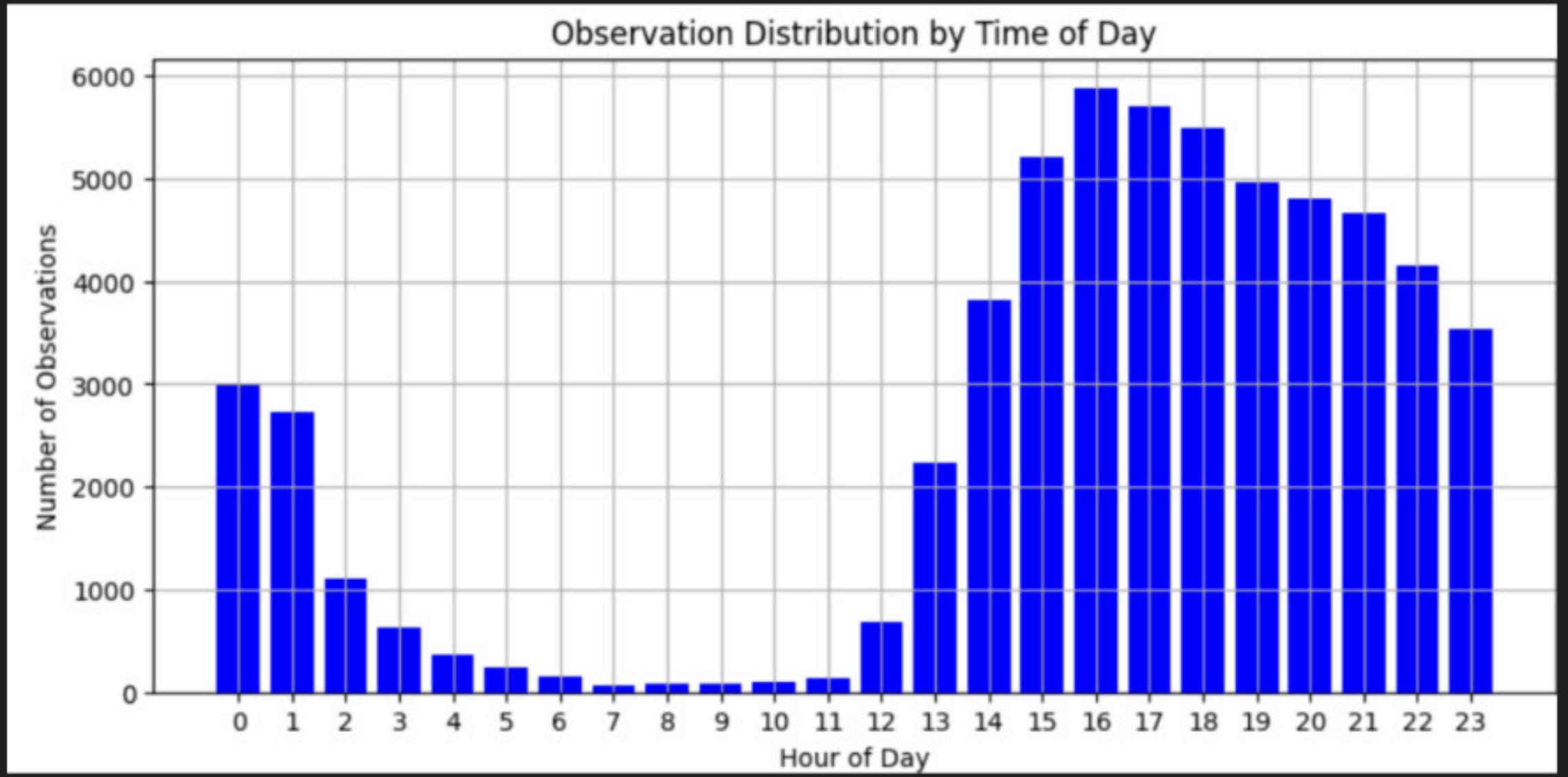
Regression Comparisons Cont.

- Decision Tree Cont.
  - RMSE - Bees, Hummingbirds
  - Max Depth
- Random Forest
  - RMSE - Bees

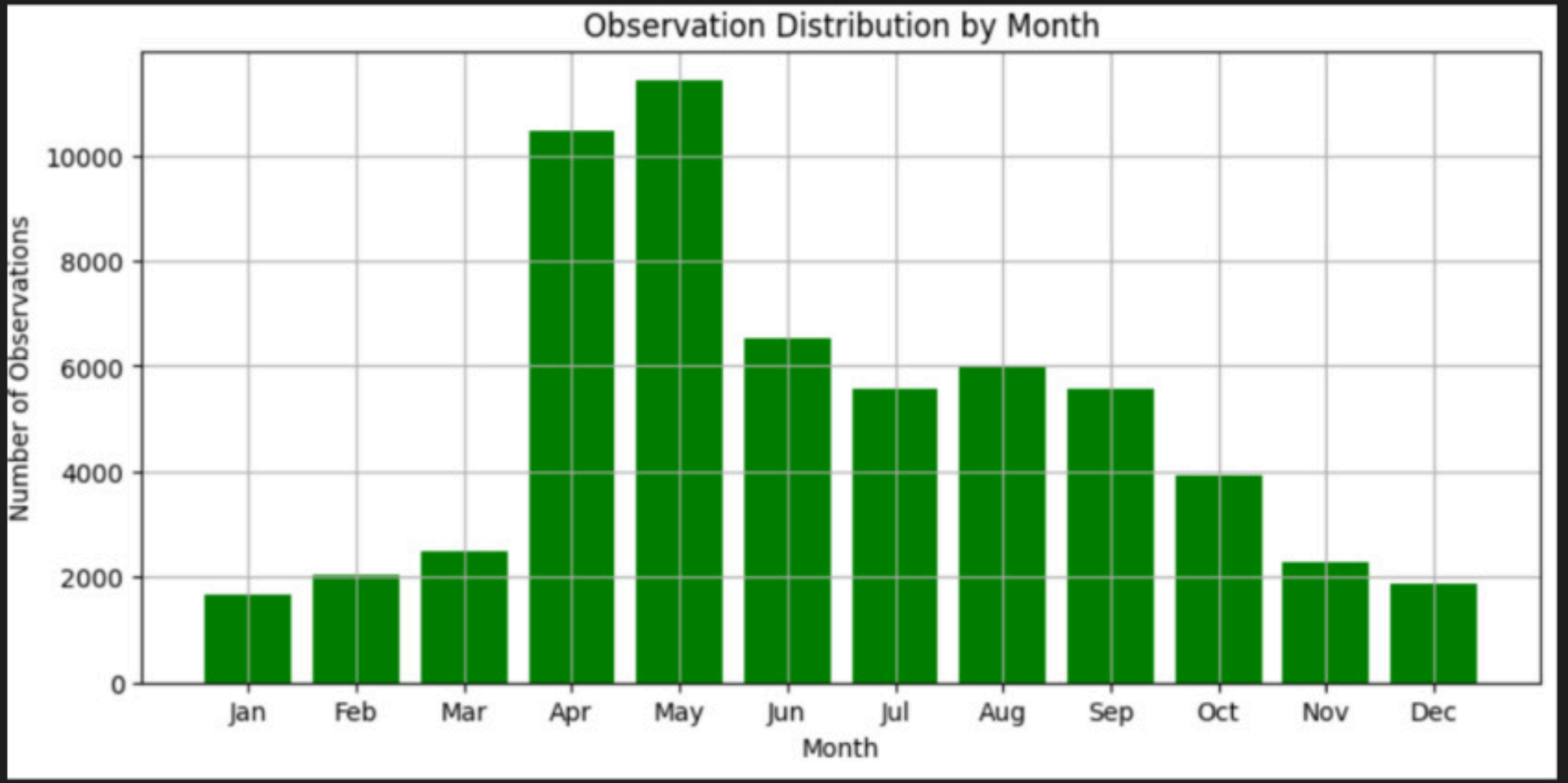


Biases/Data Struggles (UTC)

Mountain Standard Time (North America) is **7 hours** behind **UTC** universal time



Biases/Data Struggles



Future Study

- Due to time/data/budget constraints this project is limited in scope and precision, but with limited resources we have shown the incredible power of the tools of data science.
- Different data could easily be plugged into the same model to show factors in and prevalence of mental health problems, homelessness or disease to name a few.
- A few other geospatial data leads that would bring a richer dimension to the project would be climate data, terrain analysis, solar exposure, microclimates associated with structures, hydrography, etc.

Acknowledgments

We want to thank CNM staff, Job Training Albuquerque, Dekker/Perich/Sabatini, Tricore and Hewlett Packard for their support in this half year endeavour.

Special thanks to Robert Citek! We will miss the mind expanding Fridays!

We've used a constellation of tools, libraries and data sources:

DigitalOcean

jupyter

CO

PostGIS
Spatial PostgreSQL

python

PostgreSQL

matplotlib

USGS
science for a changing world

USDA

QGIS

GeoPandas

GDAL

iNaturalist

scikit learn

NumPy

DEEP DIVE
Coding
powered by CNM Ingenuity

Q&A

Nicky Hughes
nbhughes27@gmail.com
LinkedIn:

Drew Seavey
drew.seavey@gmail.com
LinkedIn:

Elizabeth Fitter
elizagethgfitter@gmail.com
LinkedIn:

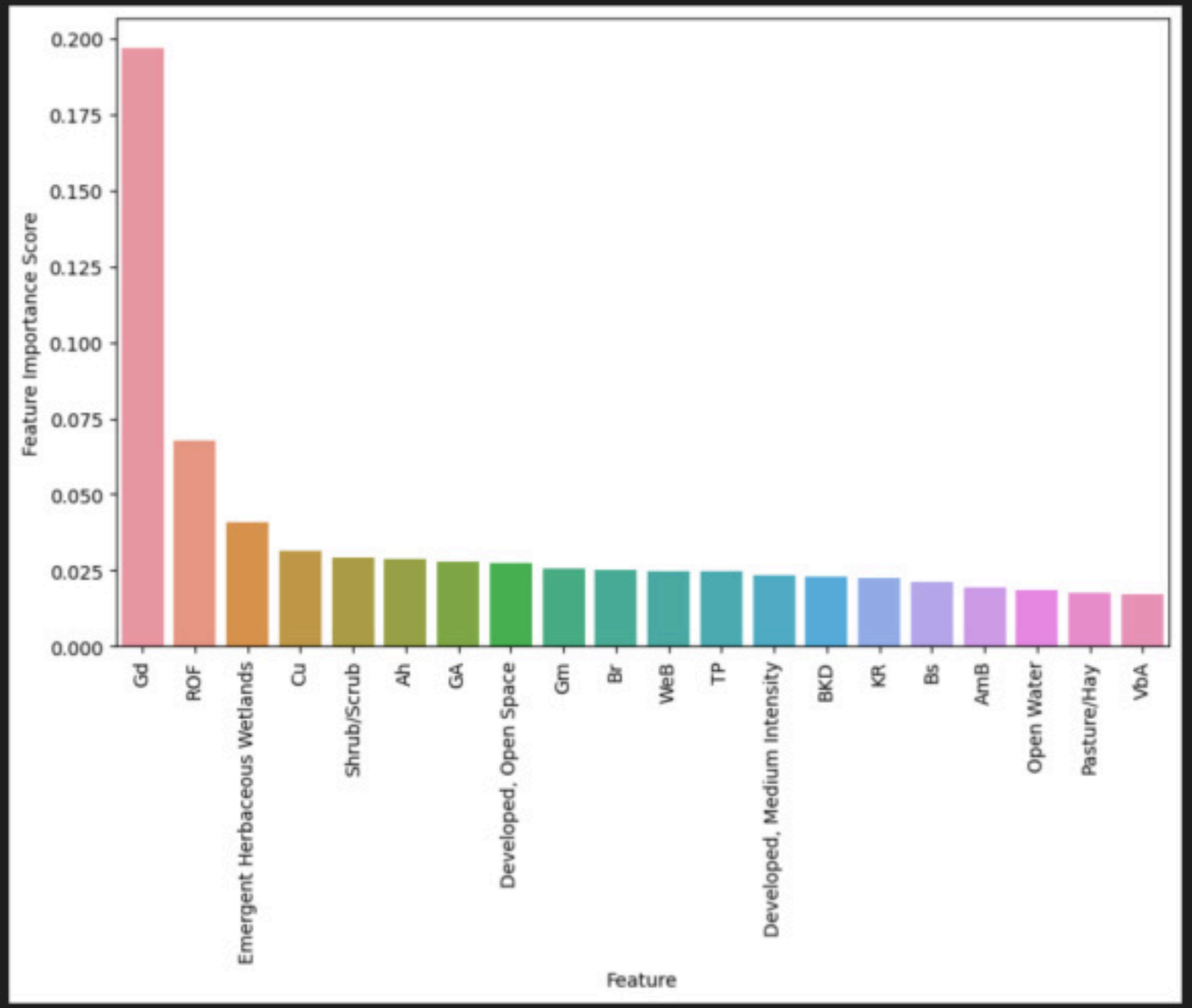
Project Repository on GitHub:

DEEP DIVE Coding
powered by CNM Ingenuity

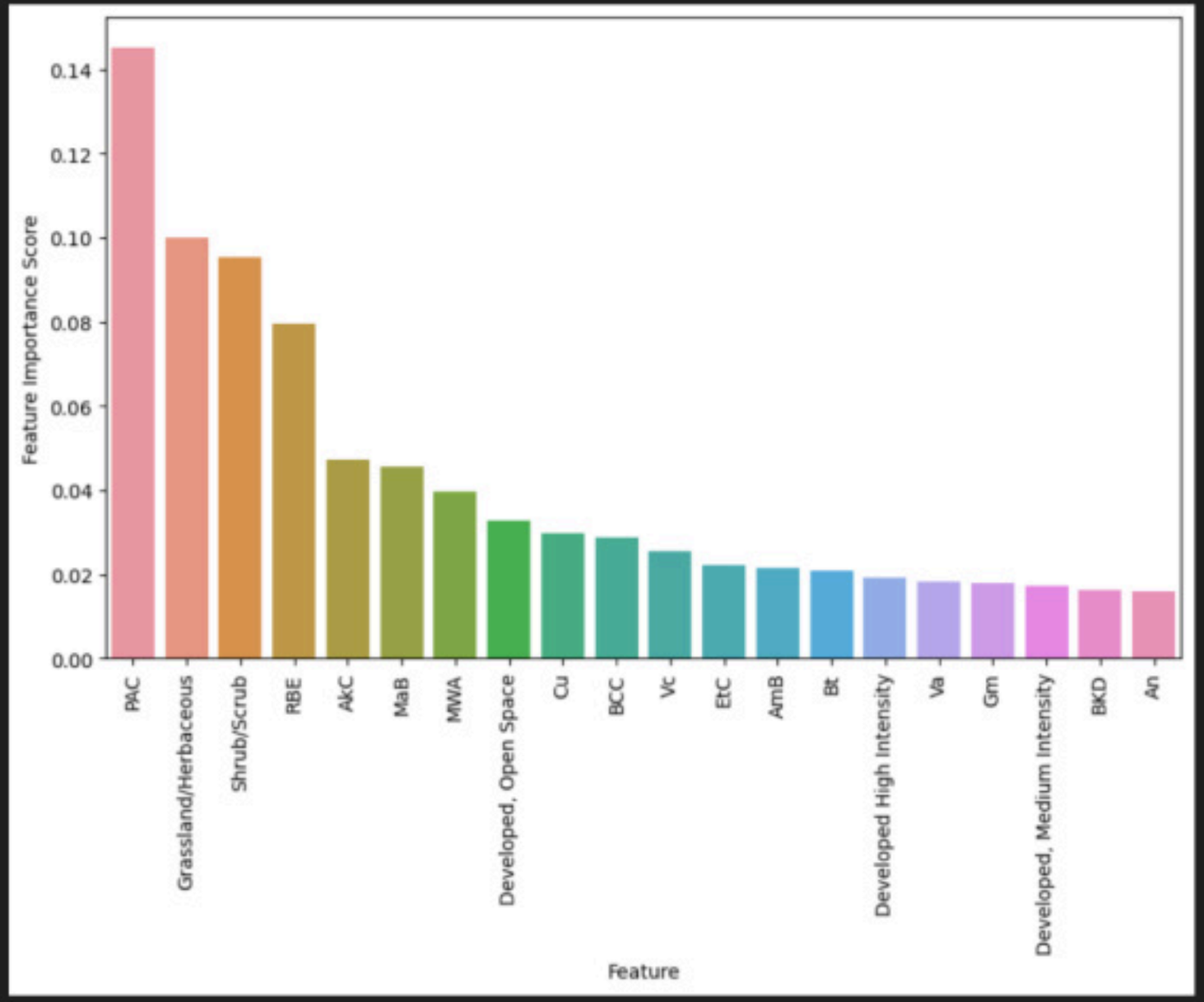
Soil Types

Symbol	Description	Slope %	Composition
AmB	Alameda sandy loam	0 to 5	sandy loam, gravelly sandy loam, very cobbly loam, bedrock
BOF	Borolls-Rock outcrop association	very steep	Very stony loam, very stony fine sandy loam, bedrock
Br	Brazito fine sandy loam		Fine sandy loam, course sand
Cu	Cut and fill land		
Gm	Glendale clay loam	0 to 1	Clay loam, silt loam, fine sand, clay
LtB	Latene sandy loam	0 to 5	Sandy loam, gravelly sandy loam
ROF	Rock outcrop-Orthids complex	40 to 80	Bedrock
VFL	Flagstone		Very flaggy

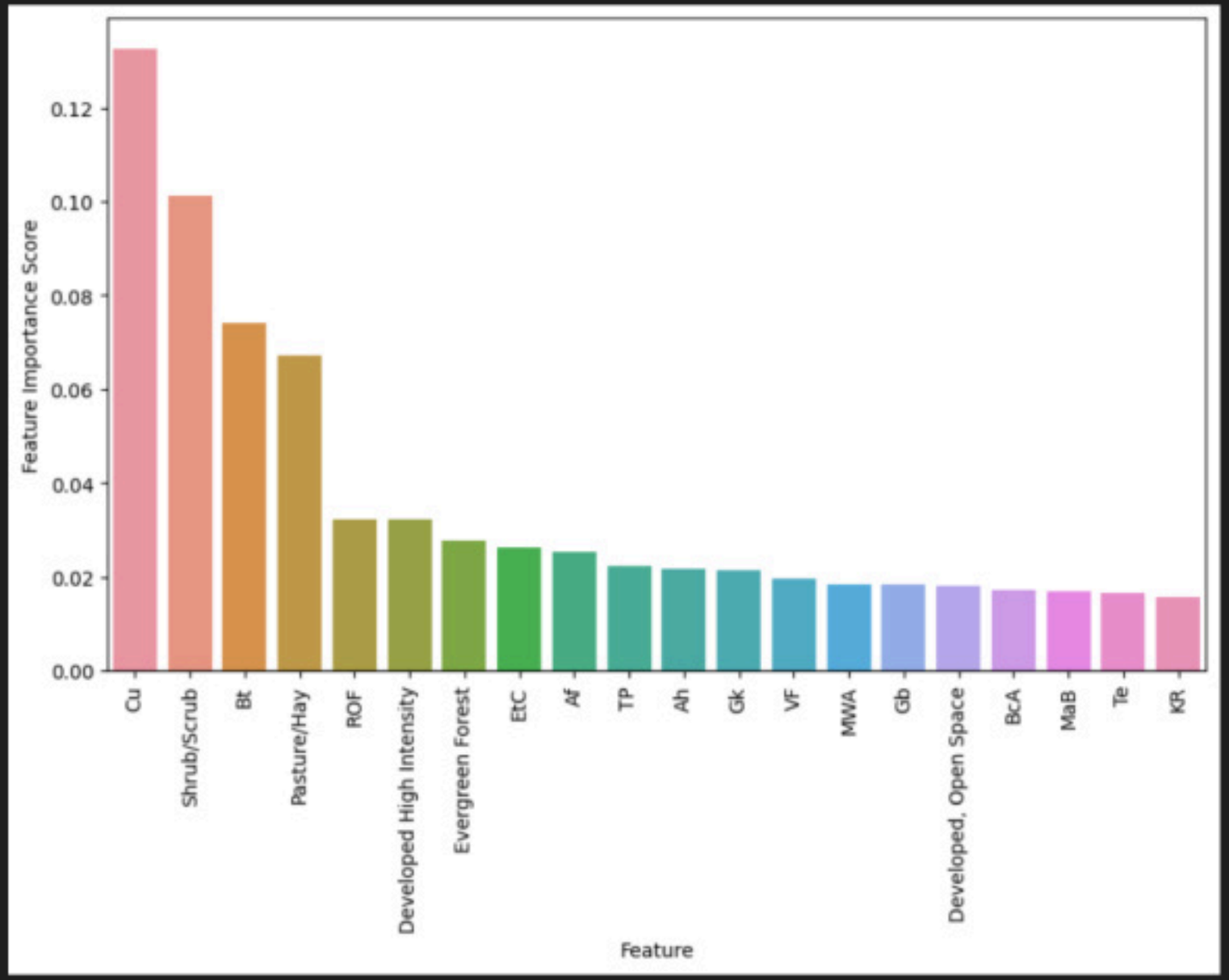
Environmental Feature Importance - Coyote



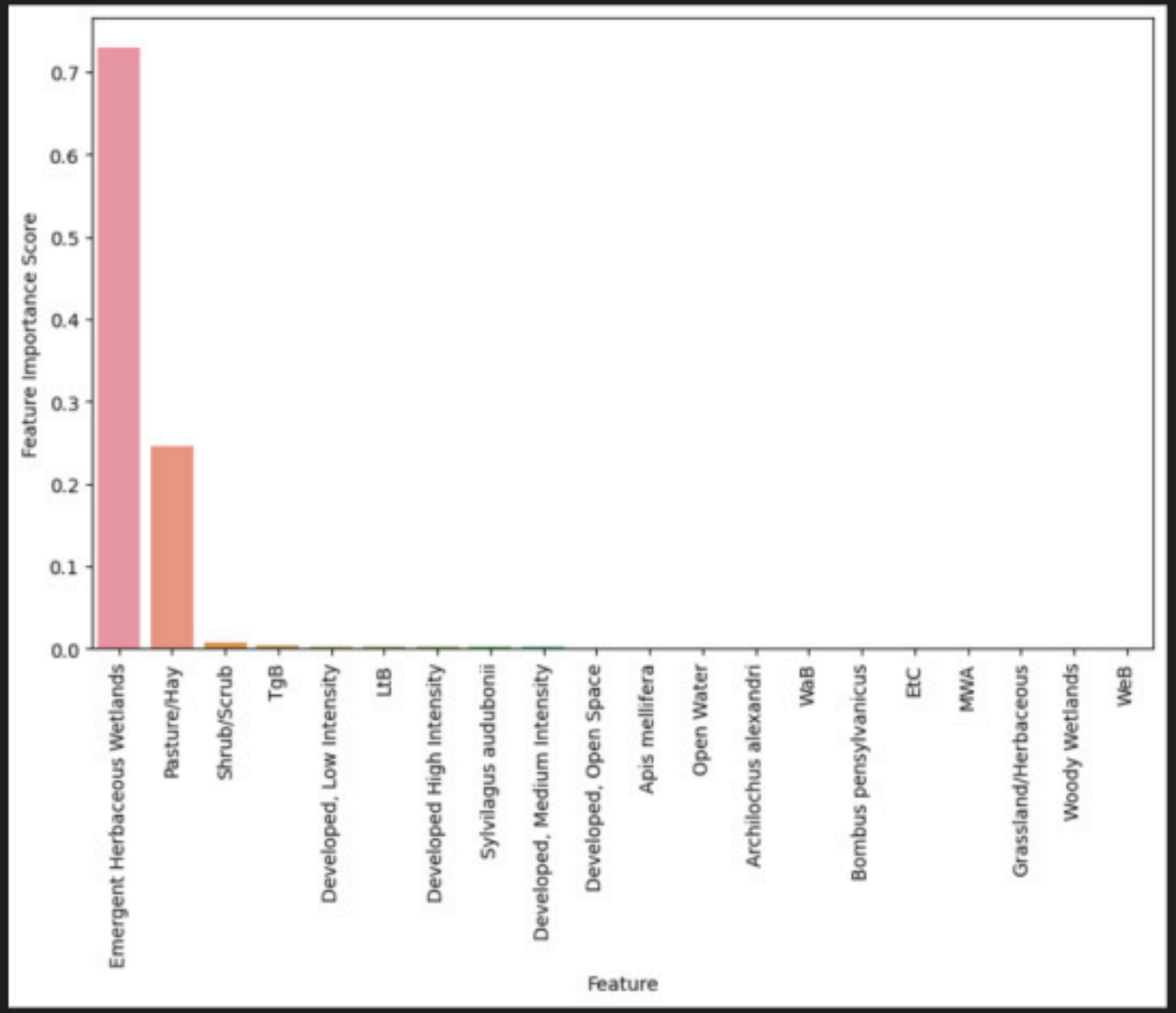
Environmental Feature Importance - Beaver



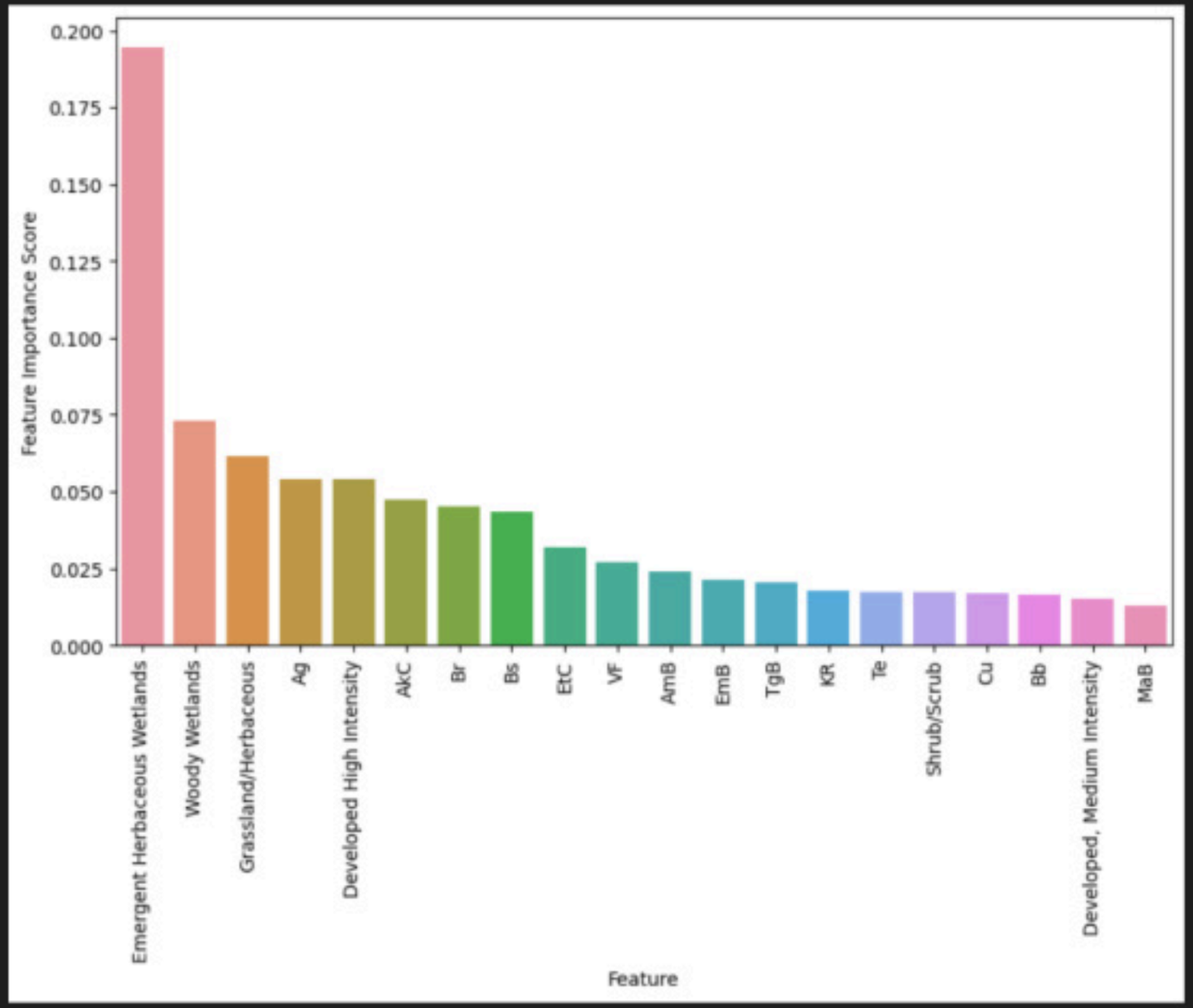
Environmental Feature Importance - Cottontail



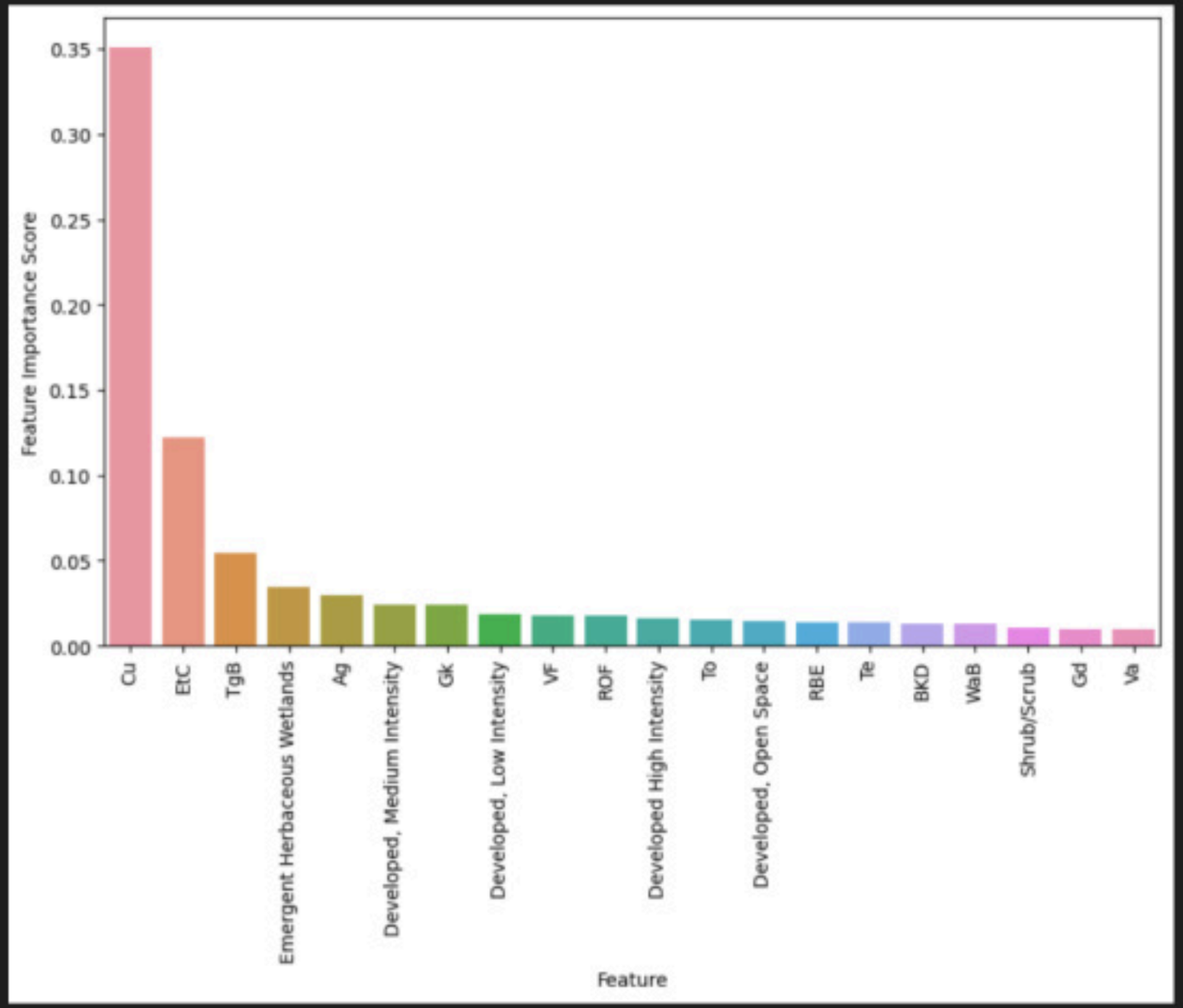
Environmental Feature Importance - Gunnison's Prairie Dog



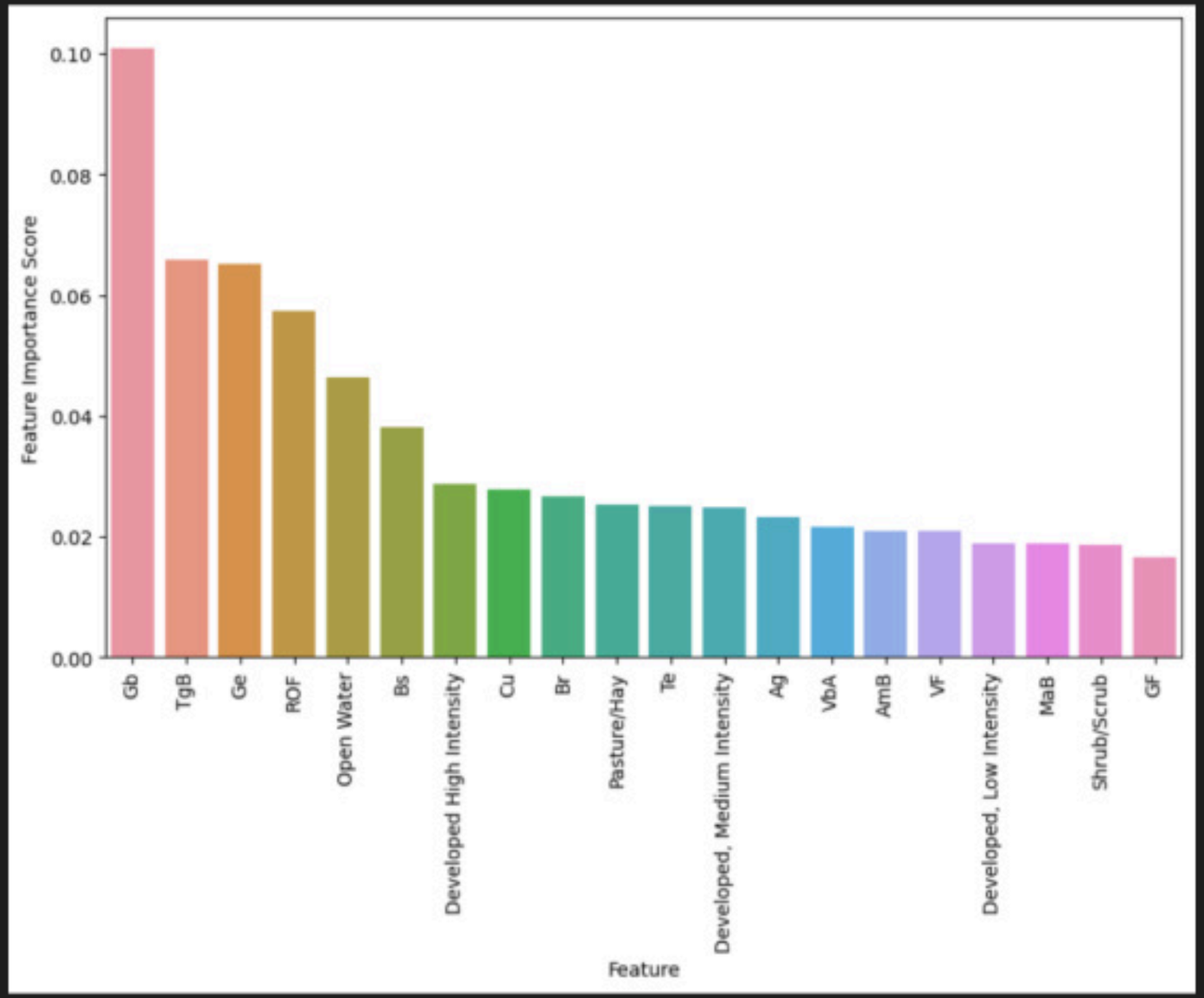
Environmental Feature Importance - Black-chinned Hummingbird



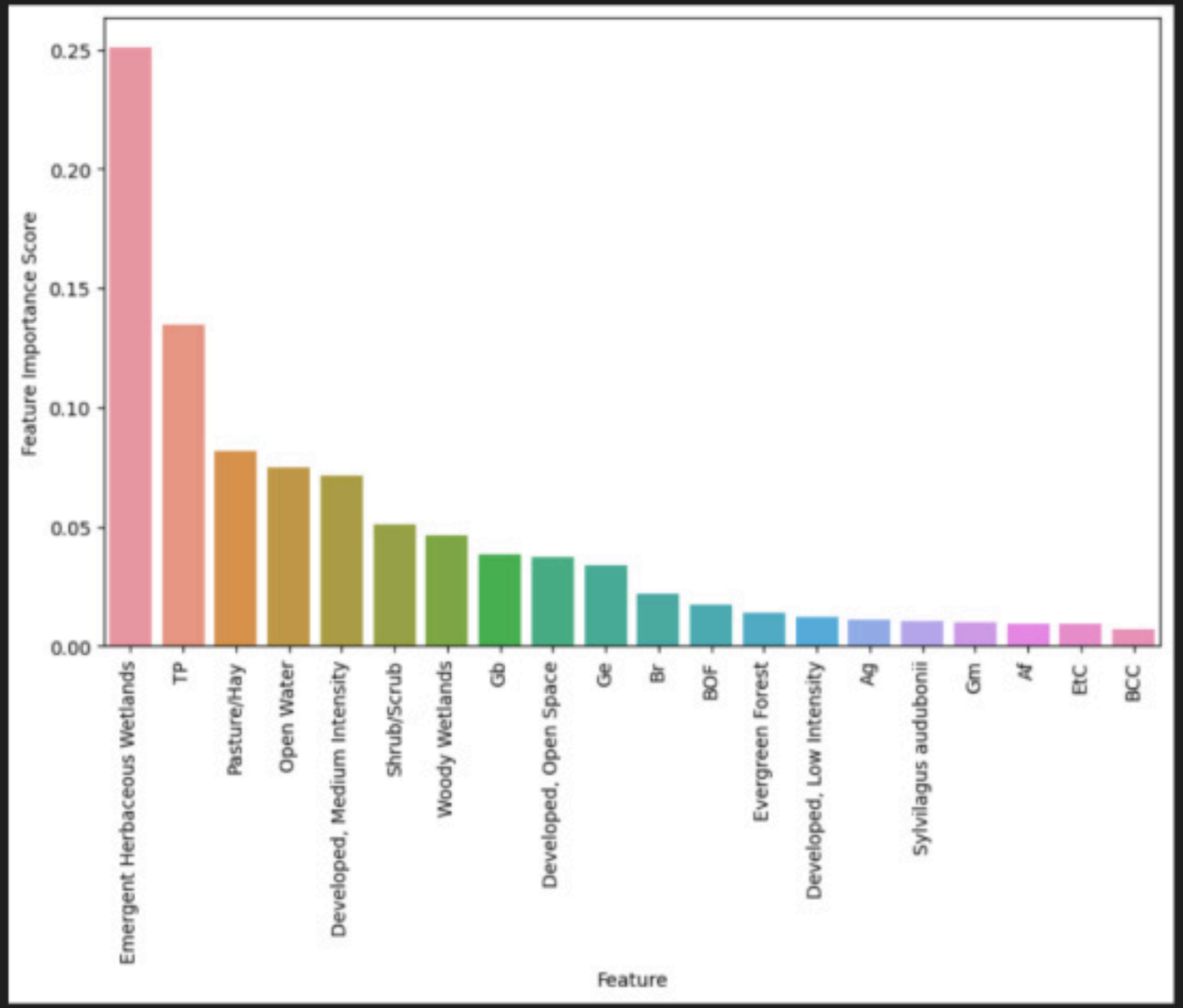
Environmental Feature Importance - Western Honey Bee



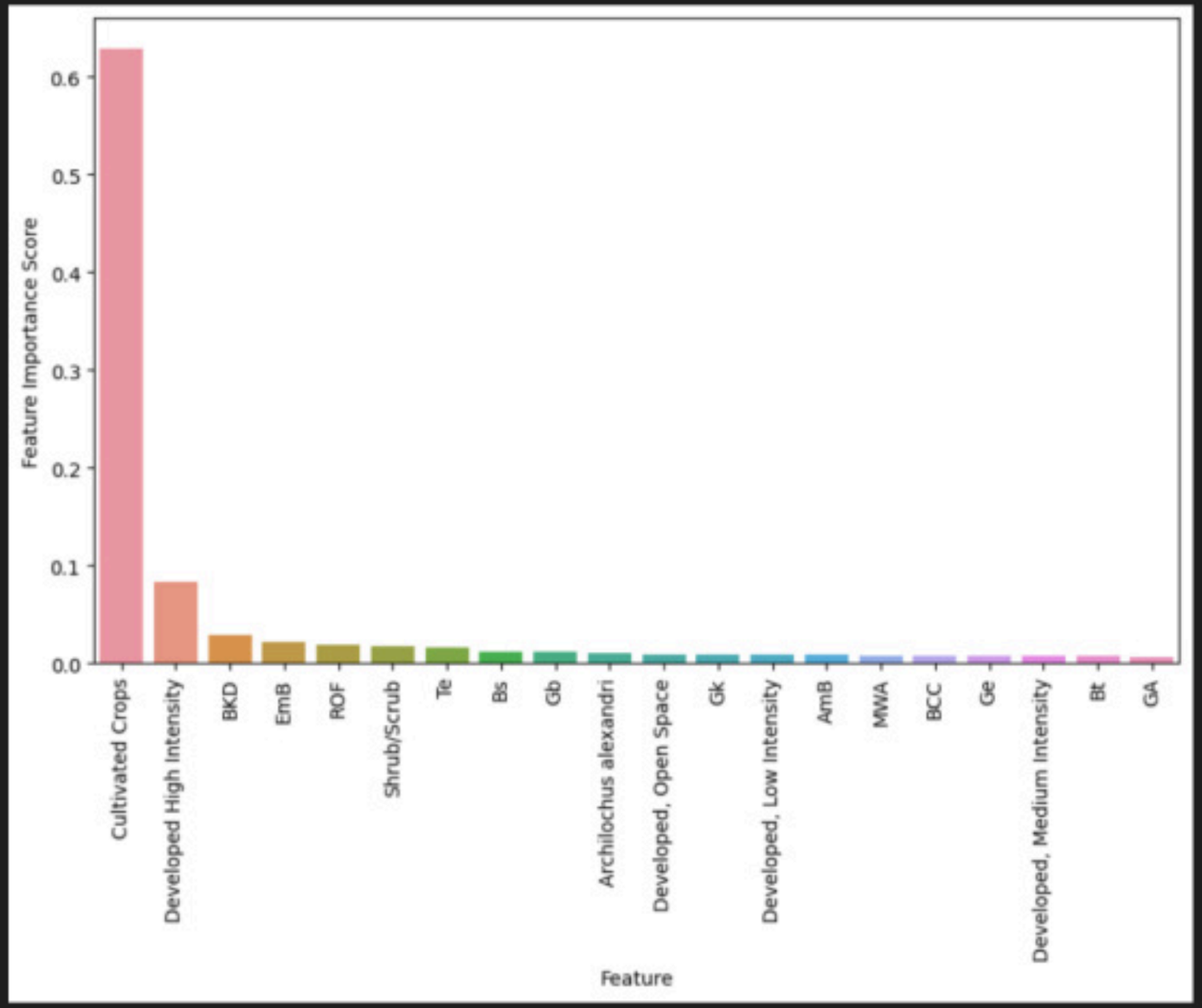
Environmental Feature Importance - American Bumblebee



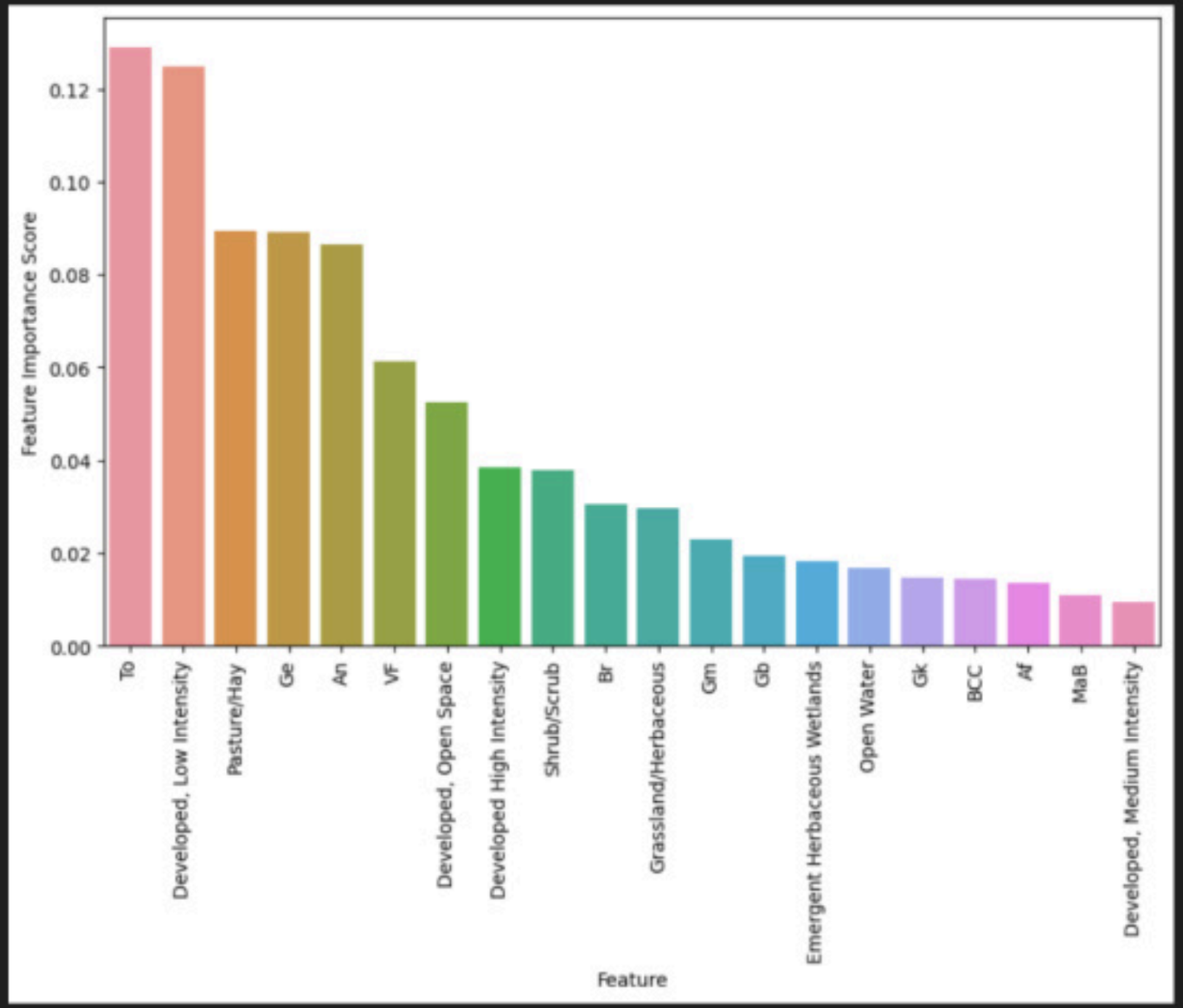
Environmental Feature Importance - Coyote Willow



Environmental Feature Importance - Chamisa



Environmental Feature Importance - Rio Grande Cottonwood



Regression Comparisons Cont.

- Logistic Regression
  - Correlation Matrix
  - Distribution of predictors
  - RMSE - bees