Predicting Diameter and Physical Harm of Asteroids using Machine Learning

Colin Campbell
Department of Computer Science
Texas State University
San Marcos, Texas, USA
Email: c_c953@txstate.edu

Leah Lewis
Department of Industrial Engineering
Texas State University
San Marcos, Texas, USA
Email: lrl68@txstate.edu

Ryan Wakabayashi
Department of Computer Science
Texas State University
San Marcos, Texas, USA
Email: rjw102@txstate.edu

Jake Worden
Department of Computer Science
Texas State University
San Marcos, Texas, USA
Email: jrw294@txstate.edu

CS 4347 Fall 2021

TEXAS STATE UNIVERSITY

Motivation

The ability to use the collected data on nearby asteroids and determine whether they present a threat to life on Earth is critical for the future of civilization. Companies like NASA and SpaceX are currently working on technologies to identify these celestial threats and their trajectory. As data pertaining to these celestial bodies becomes increasingly available, there is an imperative need to determine which attributes are key to detecting a hazardous asteroid.

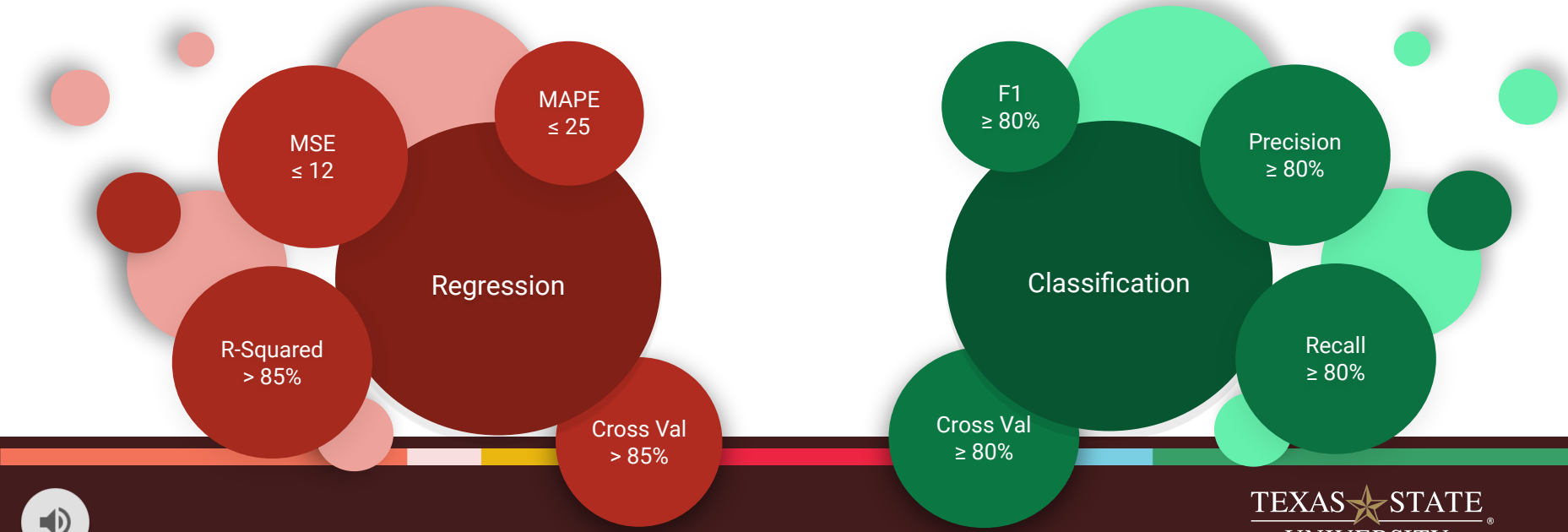
SPACEX

NASA

TEXAS STATE UNIVERSITY

Problem Statement

Given a dataset of asteroid features, can machine learning be used to predict an unknown asteroid's diameter and determine whether it's physically hazardous or not in accordance to the following success measures?



Data Management

Data Gathering

The dataset was created on behalf of NASA by the Jet Propulsion Laboratory (JPL) at California Institute of Technology's "Solar System Dynamics"(SSD) group. The dataset used here was gathered from the SSD's SBDB via the Open Asteroid Dataset challenge posted on Kaggle. The dataset itself is composed of various instances of small-bodies, along with their respective orbital elements. The dataset is comprised of 31 features and 839,714 samples.

TEXAS STATE UNIVERSITY

Data Preprocessing & EDA Overview

Classification
Regression

Visualize

- Shape
- Sum of nulls
- Correlation heatmap
- Pairplot

Encode

- Map categorical to numeric
- Label Encoder

Clean

- Drop features with >702,078 NaN
- Convert values to numeric
- Drop rest of NaN

Feature Selection

- ANOVA

Visualize

- Shape
- Sum of nulls
- Correlation heatmap
- Pairplot

Encode

- Map categorical to numeric
- Label Encoder

Clean

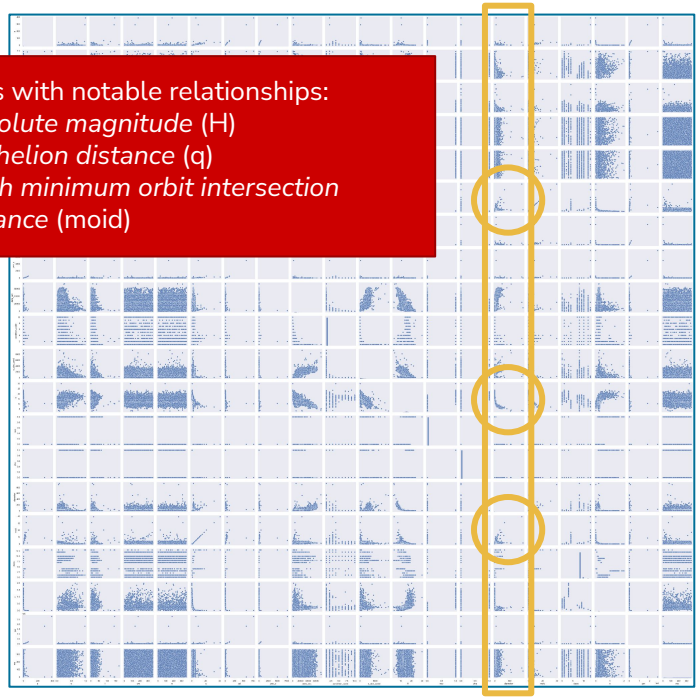
- VAE data to balance classes
- Downsample majority class
- Drop NaN like regression

Feature Selection

- ANOVA
- Principal Component Analysis

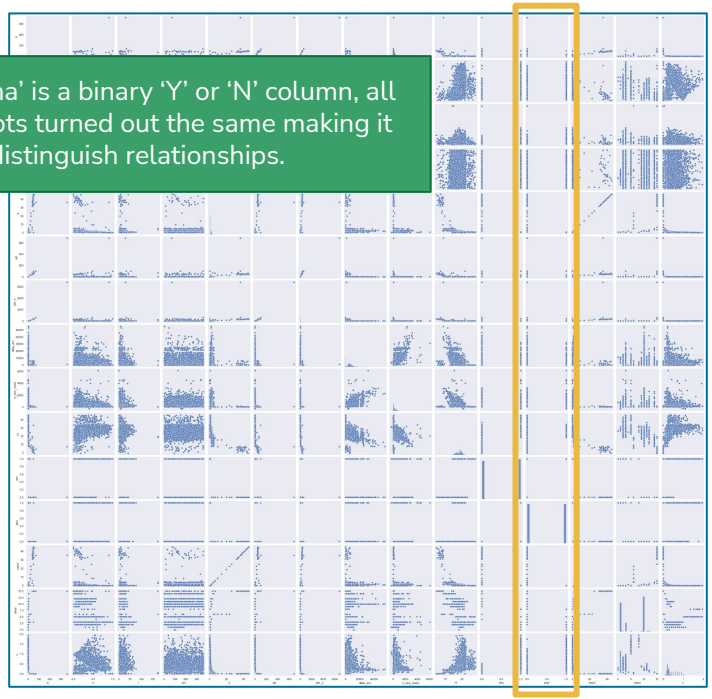
Pairplots

Classification
Regression



Features with notable relationships:

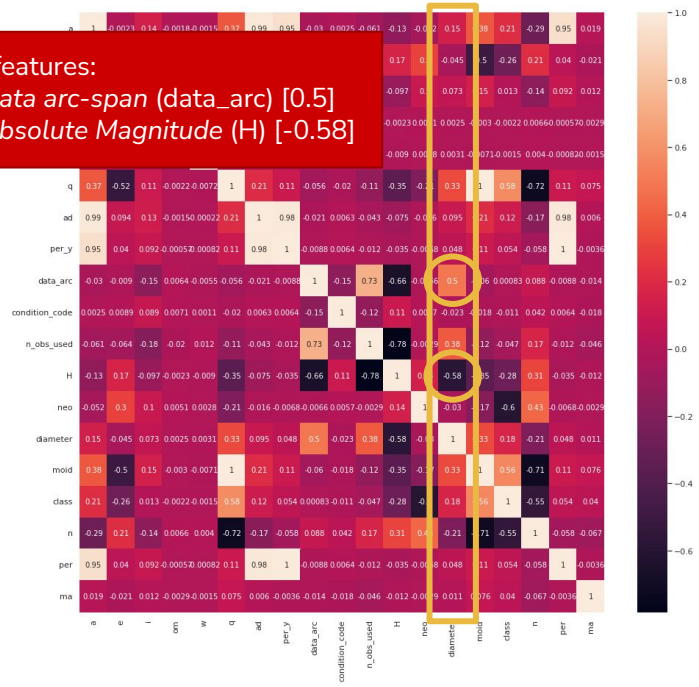
- *Absolute magnitude* (H)
- *Perihelion distance* (q)
- *Earth minimum orbit intersection distance* (moid)



Since ‘pha’ is a binary ‘Y’ or ‘N’ column, all of it’s plots turned out the same making it hard to distinguish relationships.

Heatmaps

Classification
Regression



Best features:
- *Data arc-span* (data_arc) [0.5]
- *Absolute Magnitude* (H) [-0.58]

Best features:
- *Near earth object* (neo) [0.98]
- *Asteroid orbit class* (class) [-0.96]
- *Features eccentricity* (e) [0.85]
- *Absolute magnitude* (H) [0.68]

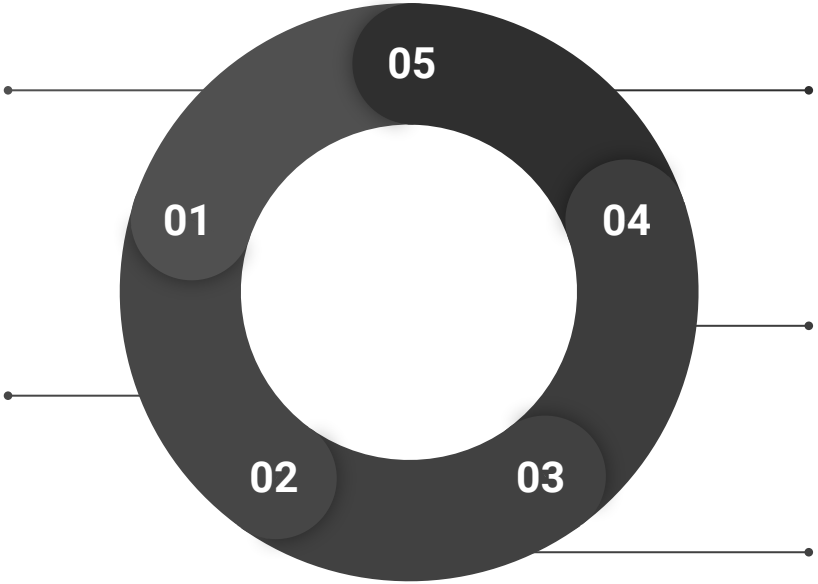
Variational Autoencoder

Initial Class Distribution

499,000 instances of class 0
1,000 instances of class 1

Variational Autoencoder

300 epochs for 6000 data samples.



Final Class Distribution

8,013 instance of class 0
8,013 instance of class 1

Rectification

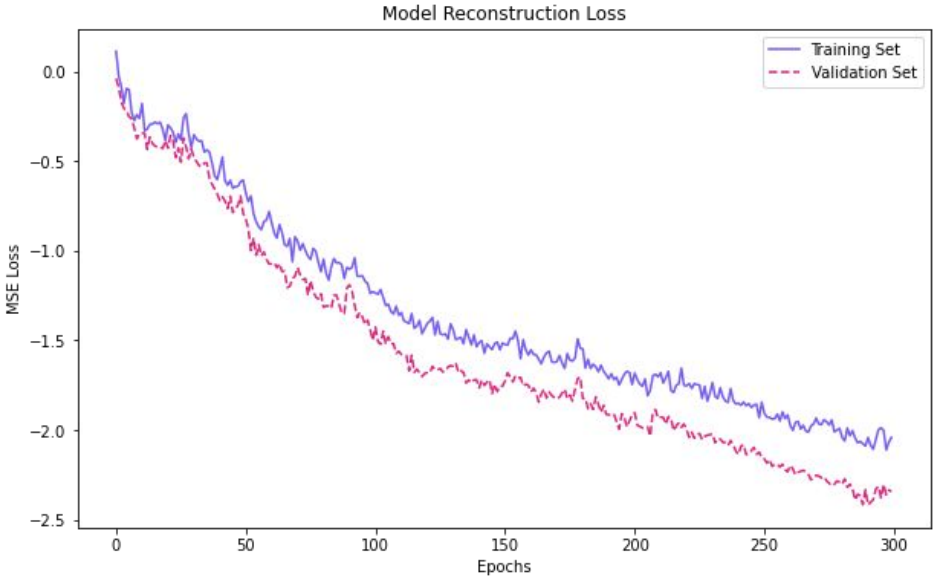
Encoded categorical data to numerical ('pha', 'neo' and 'condition_code').

Sanity Check

Sanity checked feature means and stds against original data.

Variational Autoencoder

	Original dataset		*Generated dataset*	
Feature	**Mean**	**Std**	**Mean**	**Std**
a	1.57	0.55	1.58	0.53
e	0.55	0.22	0.54	0.21
i	12.78	12.61	12.65	12.08
om	158.11	108.88	158.91	104.67
q	0.78	0.22	0.78	0.21
ad	2.36	1.07	2.41	1.04
per_y	2.05	1.08	2.10	1.05
data_arc	4969.40	4858.82	4967.25	4670.21
condition_code	1.66	1.78	1.71	1.72
n_obs_used	332.14	494.56	317.77	476.34
H	17.71	5.56	17.81	5.33
neo	1.0	0.0	0.96	0.15
pha	1.0	0.0	1.0	0.0
moid	0.11	0.28	0.104	0.27
class	1.05	0.38	1.05	0.37
n	0.63	0.32	0.61	0.32



Feature Set Selection

Classification
Regression

Orbital period year (per_y)
Orbital period day (per)
Aphelion distance (ad)
Semi-major axis (a)
Absolute magnitude (H)
Perihelion distance (q)
Earth min. orbit intersection distance (moid)
Near earth object (neo)
Mean motion deg (n)
Asteroid orbit class (class)

Near earth object (neo)
Class (class)
Eccentricity (e)
Absolute magnitude (H)
Mean motion (n)
Perihelion distance (q)
Earth min. orbit intersection distance (moid)
Inclination with respect to x-y ecliptic plane (i)
Longitude of the ascending node (om)
Number of observations (n_obs_used)

Machine Learning Approaches

Regression Models to Predict Diameter

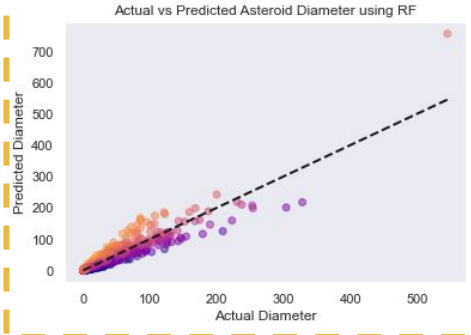
Random Forest Regressor
- Supervised ensemble method.
- Chosen for its tendency to create high performing predictions.
- Can be difficult to interpret results.

K Nearest Neighbors Regressor
- Approximates the association of a continuous target based on the samples surrounding the data point.

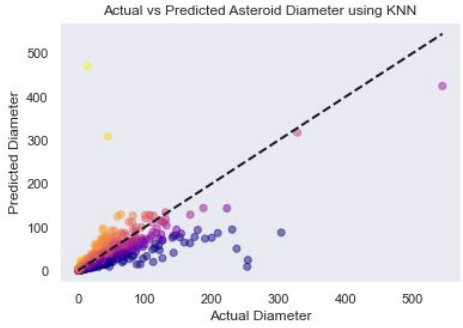
Support Vector Regressor
- Can model both linear and non-linear relationships between variables.
- Allows us to choose tolerances for errors.

Gradient Boosting Regressor
- Ensemble method
- Allows for the optimization of arbitrary differentiable loss functions.

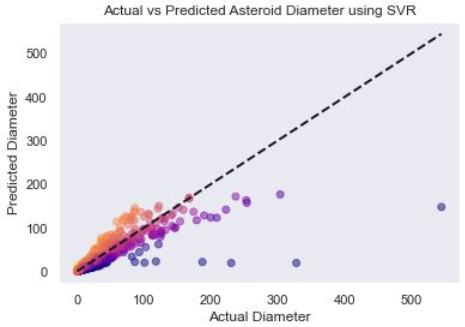
Regression Results - R^2 graphs



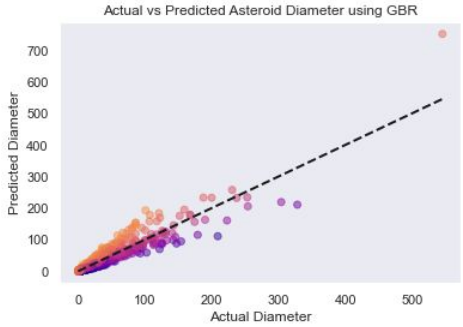
R-Squared	88.89 %
MSE	10.77
MAPE	22.17 %
10-Fold CV	87.73 %



R-Squared	63.67 %
MSE	35.48
MAPE	25.68 %
10-Fold CV	63.47 %



R-Squared	76.44 %
MSE	23.01
MAPE	26.09 %
10-Fold CV	76.12 %



R-Squared	88.56 %
MSE	11.17
MAPE	23.53 %
10-Fold CV	88.88 %

TEXAS STATE UNIVERSITY

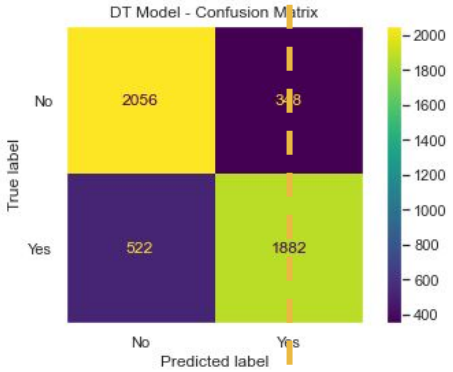
Classification Models to Predict Physically Hazardous Asteroids

Logistic Regression
- Simple classification method, used as a base point to compare other methods.
- Simple to create and execute.

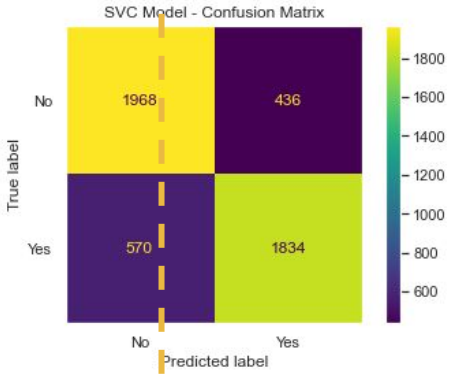
Support Vector Classification
- Works relatively well when there is a clear margin of separation between classes.
- Typically perform very well on linear and non-linear data.

Decision Tree Classifier
- Supervised learning method that creates a decision tree to make classifications.
- Well known and trusted to make high performing models.

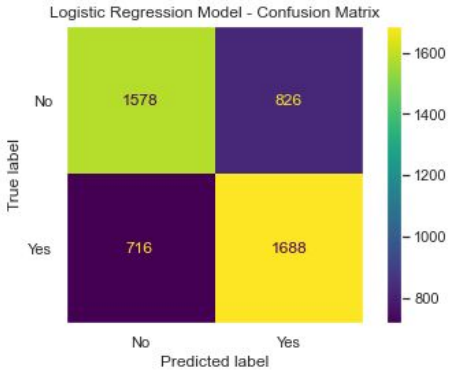
Classification Results - Confusion Matrices



Measure	Class 0	Class 1
Precision	86 %	78 %
Recall	80 %	84 %
F1	83 %	81 %
5-Fold CV	82.6 %	



Measure	Class 0	Class 1
Precision	82 %	76 %
Recall	78 %	81 %
F1	80 %	78 %
5-Fold CV	73.9 %	

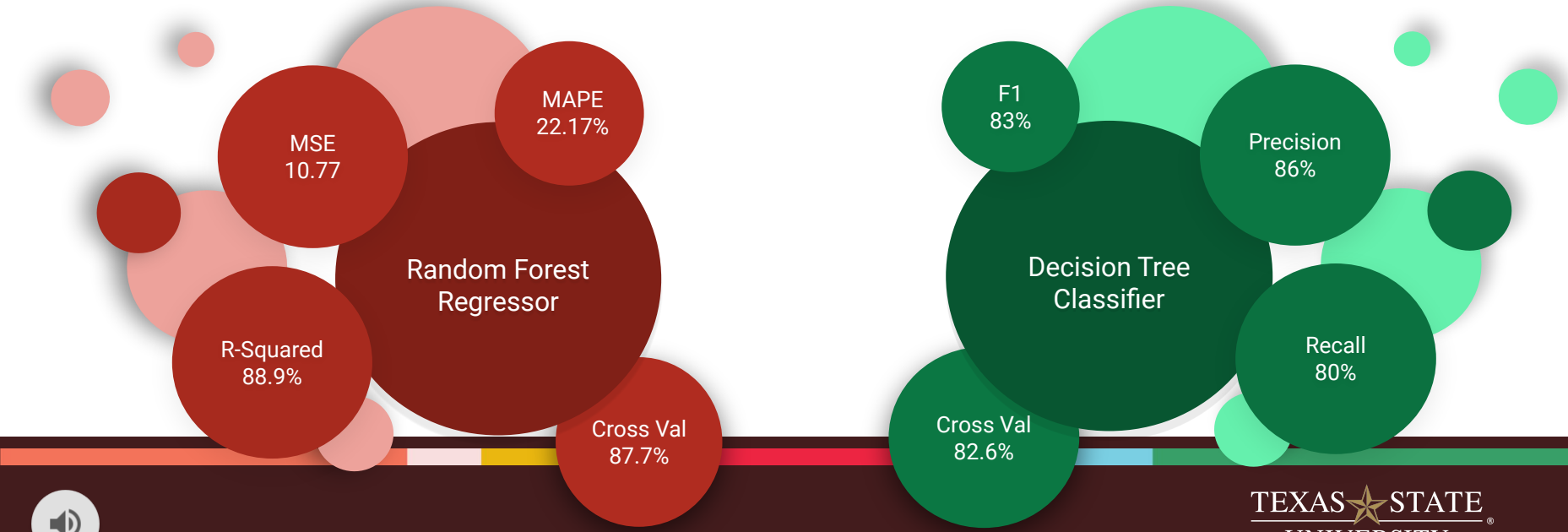


Measure	Class 0	Class 1
Precision	66 %	70 %
Recall	69 %	67 %
F1	67 %	69 %
5-Fold CV	67.1 %	

Conclusions

Conclusions

Given a dataset of asteroid features, machine learning can be used to predict an unknown asteroid's diameter with 88.9% accuracy and a precision of 86% when classifying whether it is physically hazardous or not.



TEXAS STATE UNIVERSITY

Future Work

- Encode more features during data management and preprocessing to retain more information on the dataset.
- Explore using diameter as a possible feature for classification.
- As more data is acquired, modeling approaches should improve.

References

https://scikit-learn.org/stable/

https://www.tensorflow.org/api_docs/python/tf

https://seaborn.pydata.org/

https://matplotlib.org/stable/index.html

https://numpy.org/

https://pandas.pydata.org/docs/reference/index.html

Thank you