EXOPLANETS IDENTIFICATION PROJECT

Project group 193

Hadrien Lacroix

OBJECTIVE

Identify stars that could be orbited by an exoplanet.

Exoplanets are planets outside of our solar system. Identifying exoplanets is the first step to potentially finding life outside of our solar system, potentially an intelligent form, and potentially one that can release signals into space).

DATA

Training set

***5,087** star observations, **1** label (0 or 1)**, 3197** luminosity readings for each star*

Testing set

***570** star observations, **1** label (0 or 1)**, 3197** luminosity readings for each star*

No outliers or missing values detected

HOW?

This is a classification problem: we need to classify stars as being orbited by an exoplanet (1) or not (0).

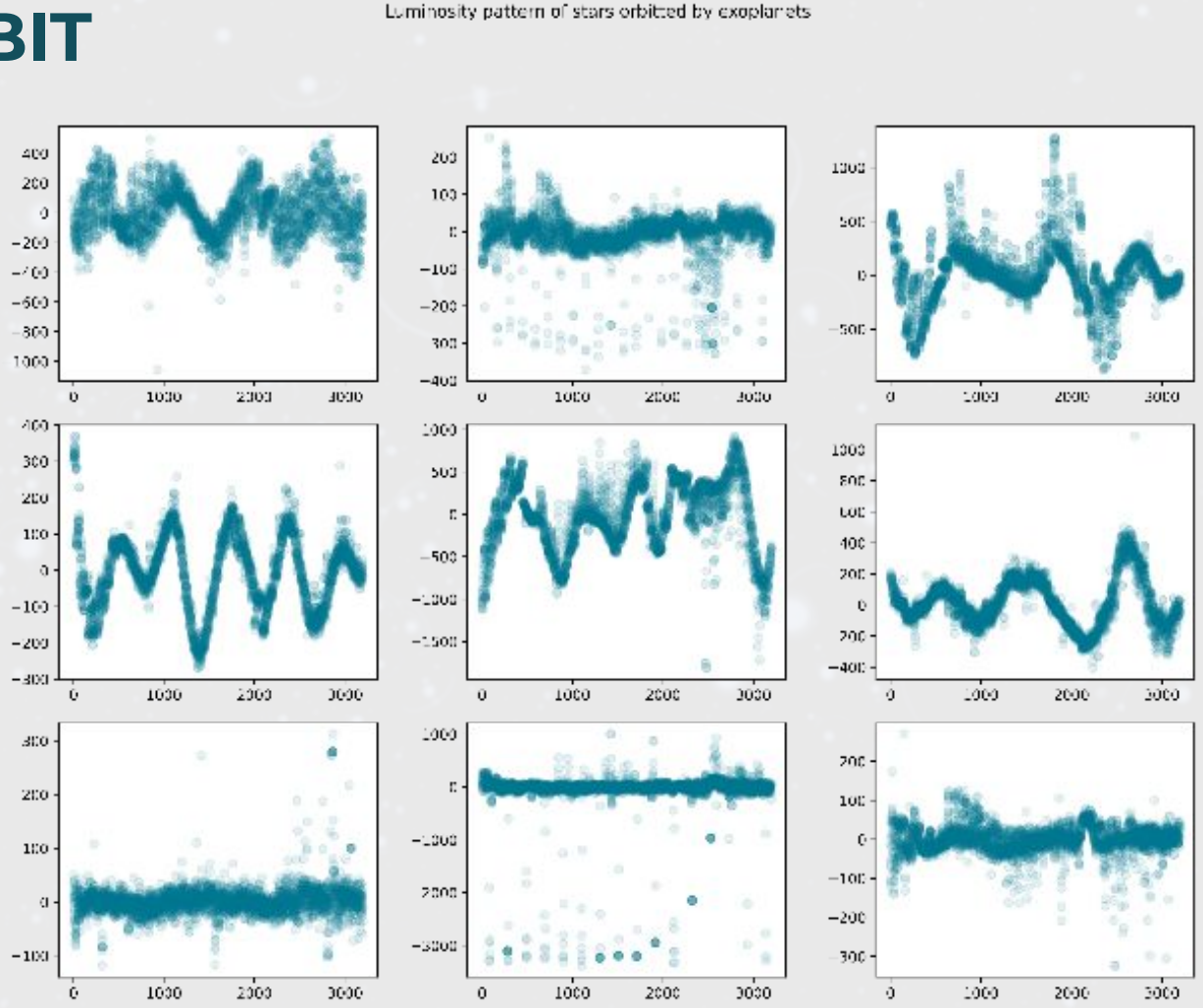
Exoplanets are detected by observing the luminosity of stars and seeing whether or not they dim at a regular interval. If so, it's possible that this dim is caused by a planet orbiting around it.

VISUALIZING ORBIT

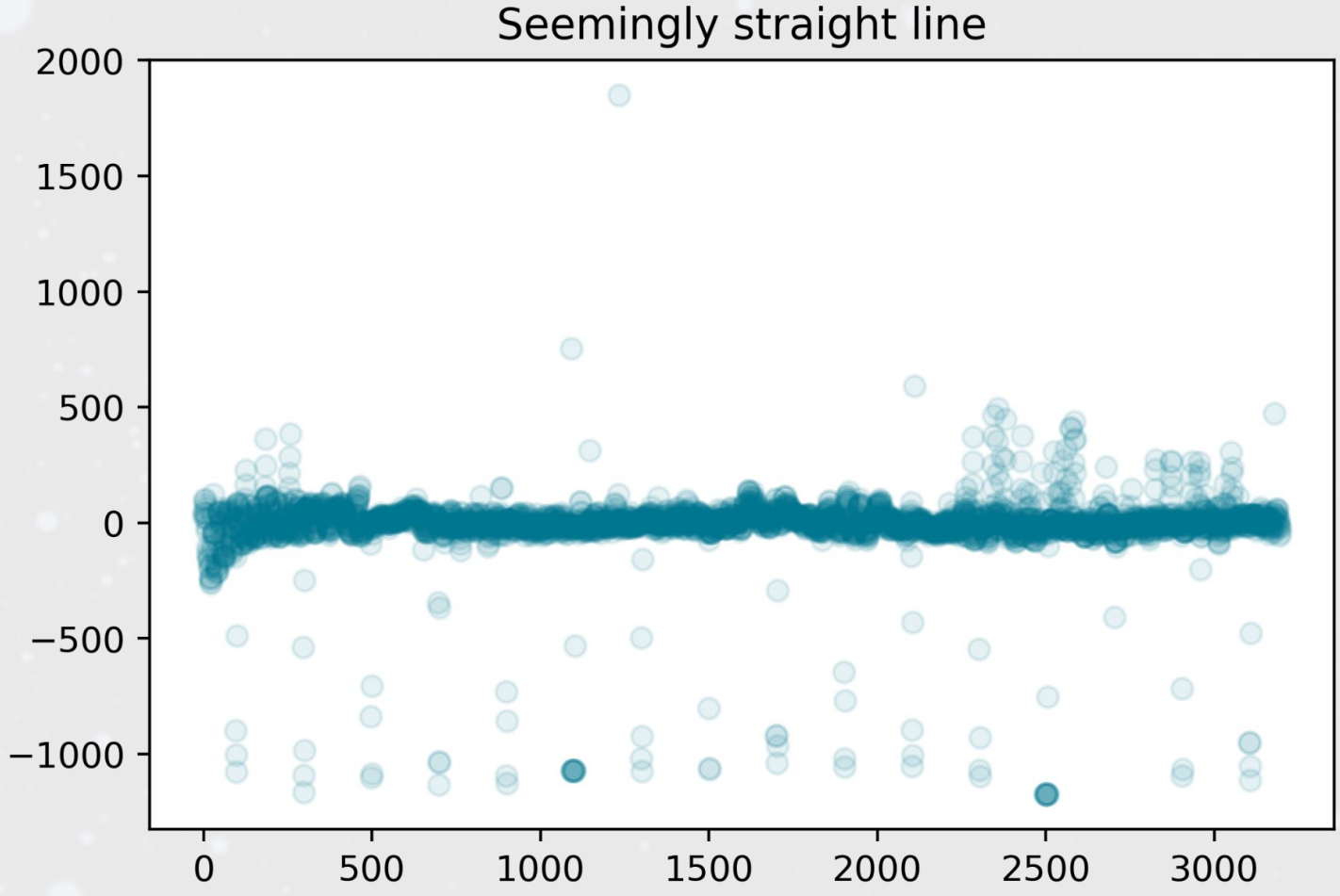
Sinusoidal movement

luminosity dims as a planet moves in front of the star and increases when it passes behind.

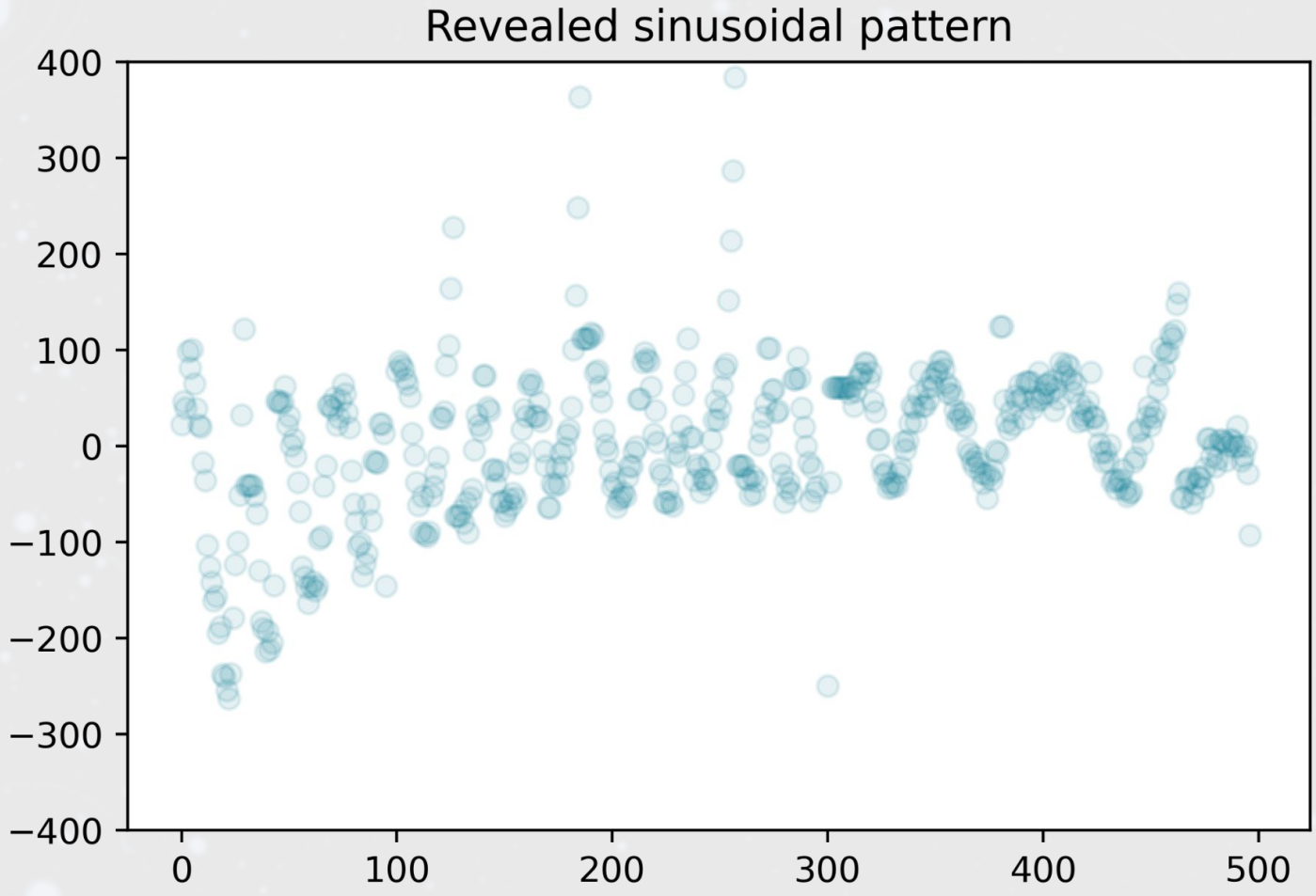
Similar highs and lows.



VISUALIZING ORBIT



VISUALIZING ORBIT



CHALLENGE

Imbalance

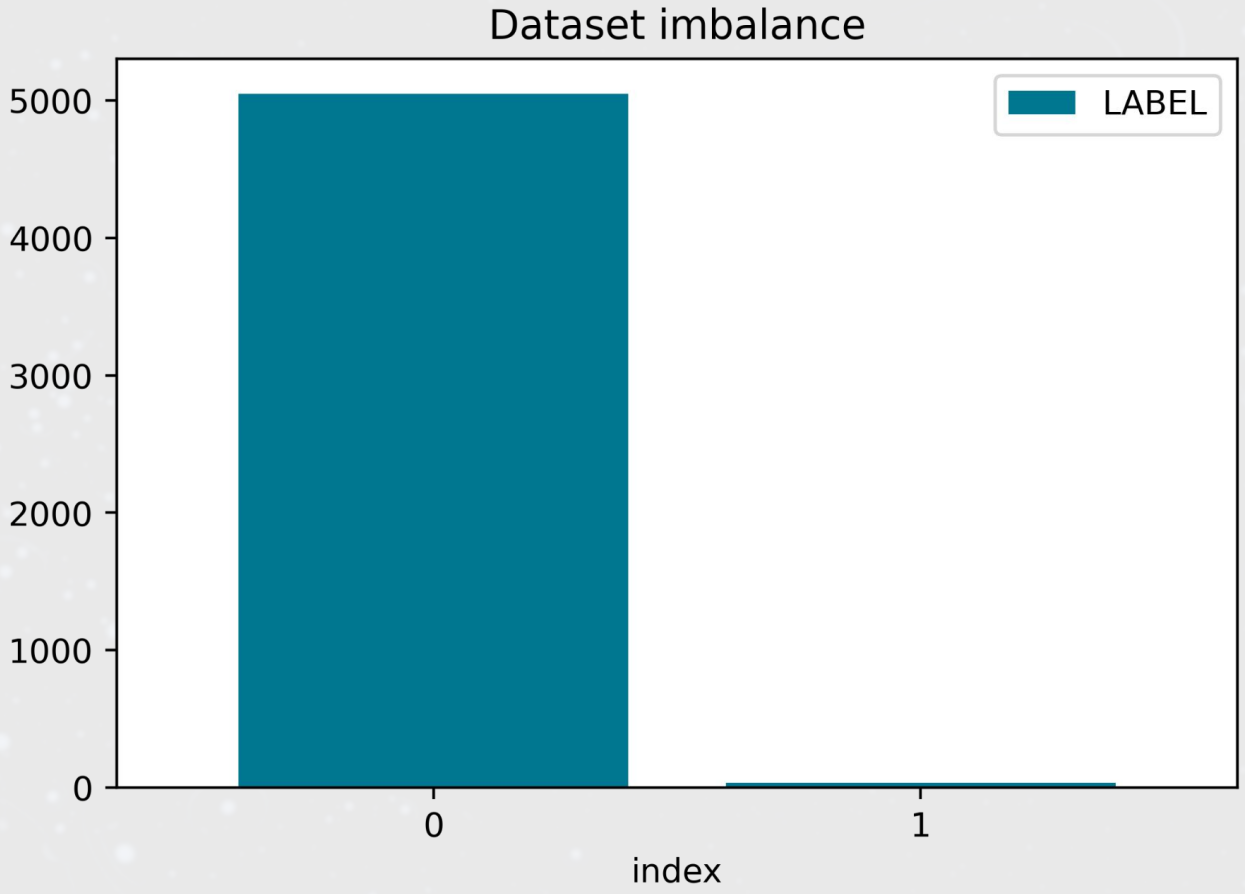
*Exoplanets are **rare**.*

*In our **training** dataset, we have **5,087** observations but only **37** exoplanets.*

*In our **testing** dataset, we have **570** observations but only **5** exoplanets.*

*We could just predict that no star is orbited by an exoplanet and reach a **99.12% accuracy**, while still completely **failing** to meet our objective (**to identify exoplanets**).*

VISUALIZING IMBALANCE



SOLUTION

SMOTE

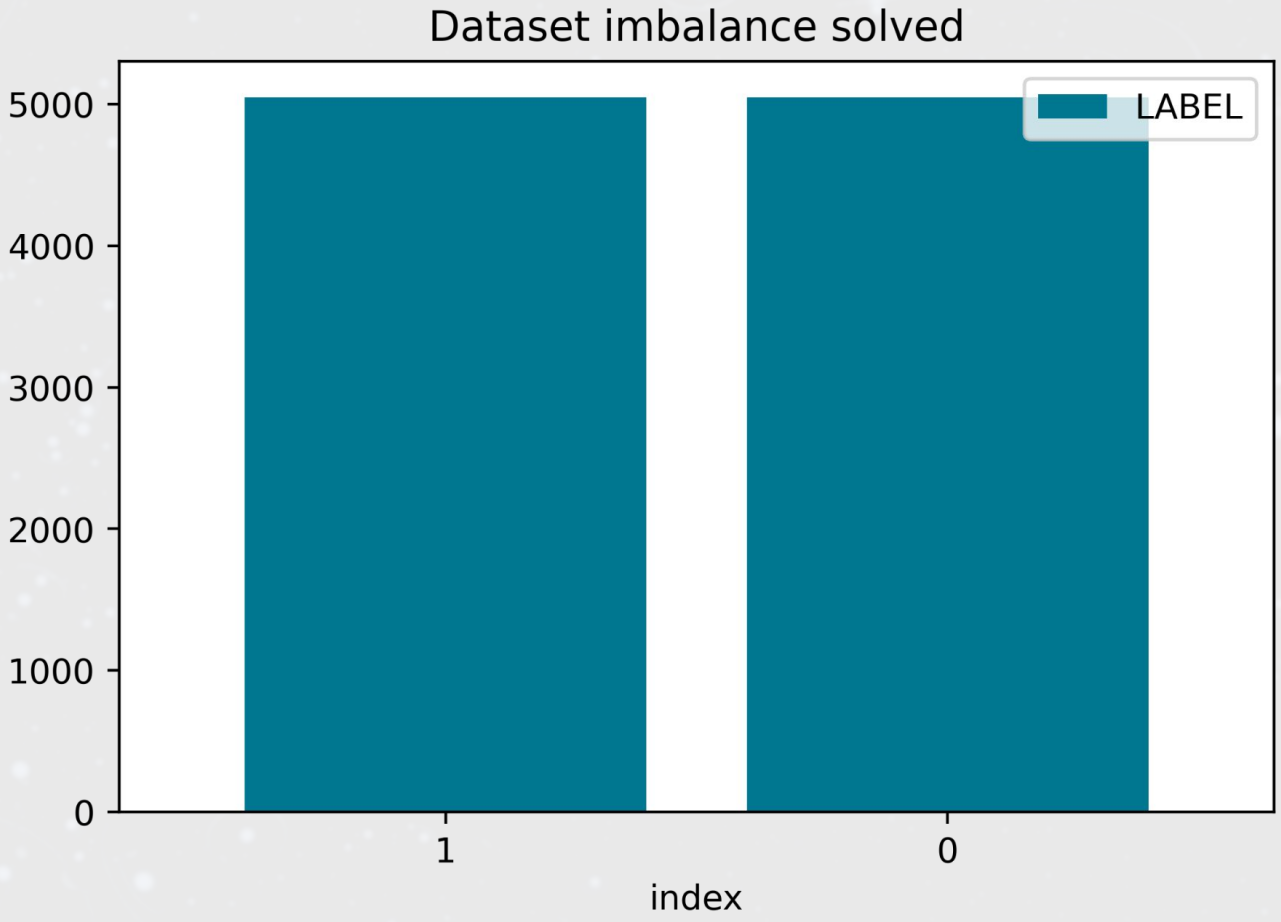
(Synthetic Minority Oversampling Technique)

*Artificially **increase the number of positive observations** in the dataset*

*by **taking samples** of the feature space for each target class and its **nearest neighbors**,*

*and **generating new examples** that combine features of the target case with features of its neighbors.*

SOLVED IMBALANCE



APPROPRIATE METRIC

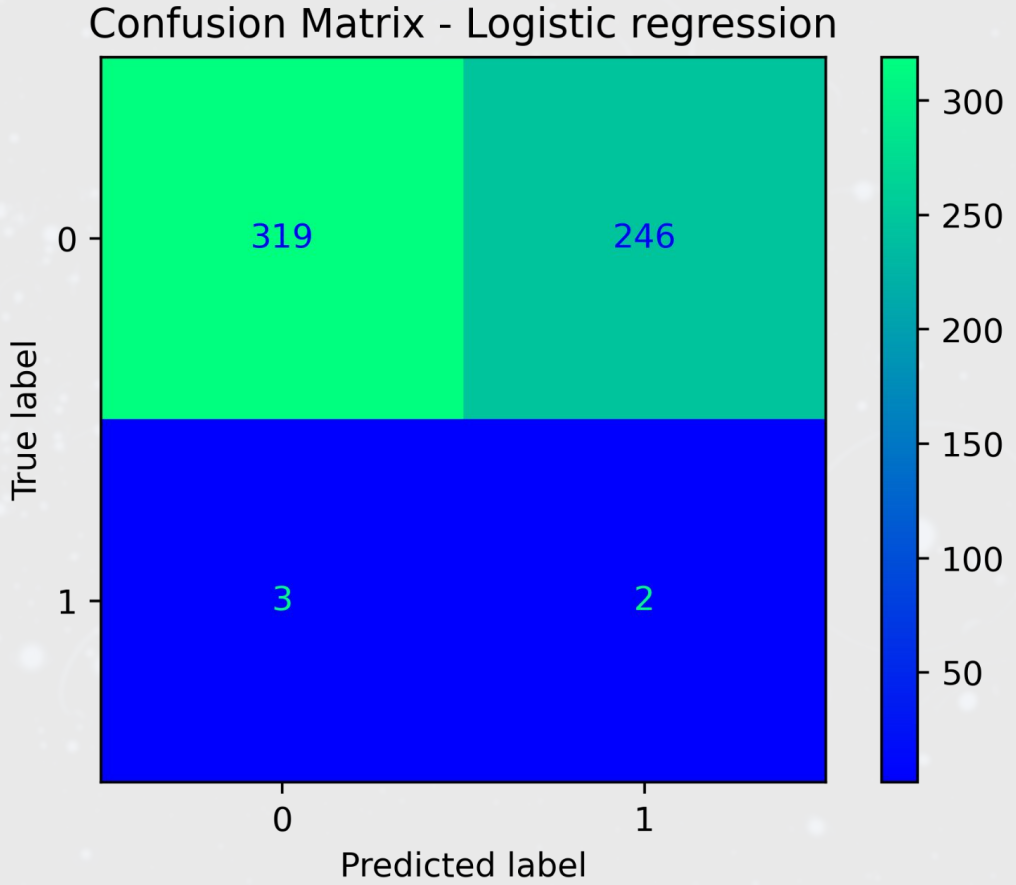
Precision

$$\frac{\textit{True Positives}}{\textit{True Positives + False Positives}}$$

MODELING - LOGISTIC REGRESSION

Before SMOTE

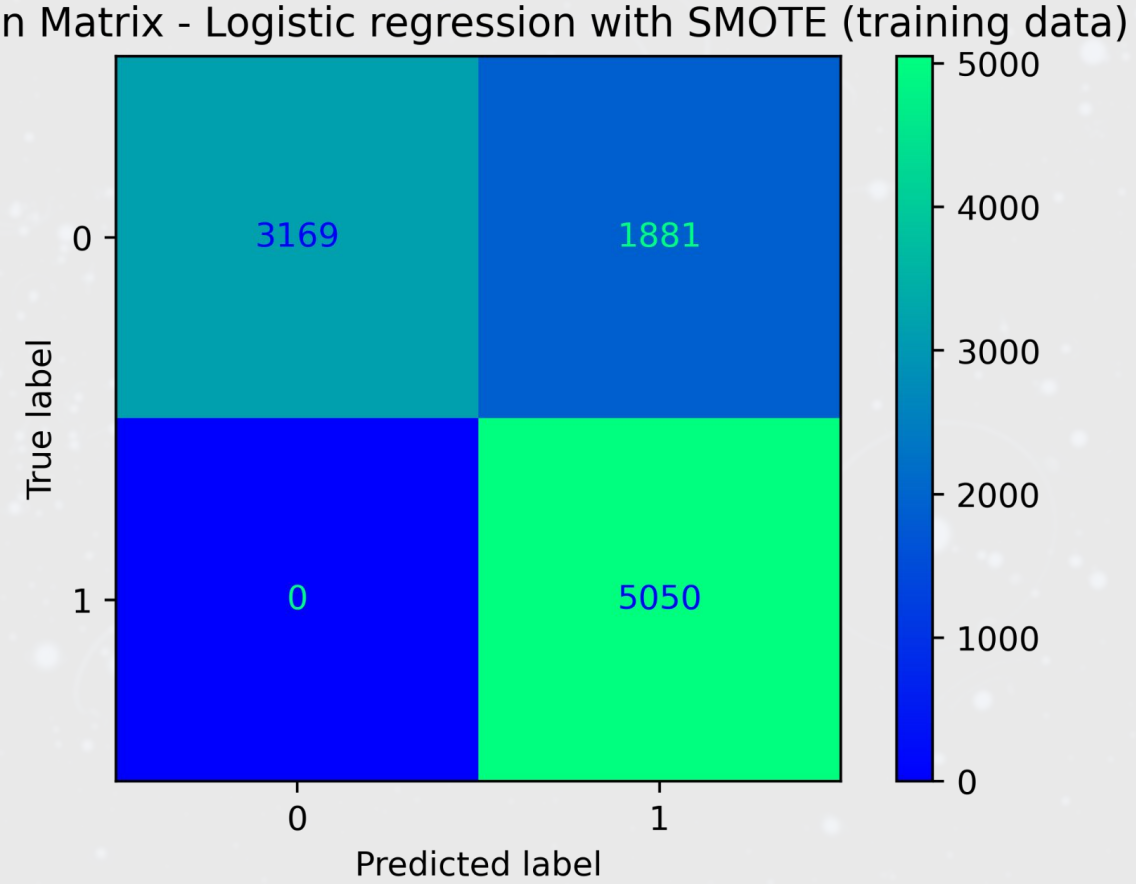
Precision: 0.4



MODELING - LOGISTIC REGRESSION

On SMOTE data

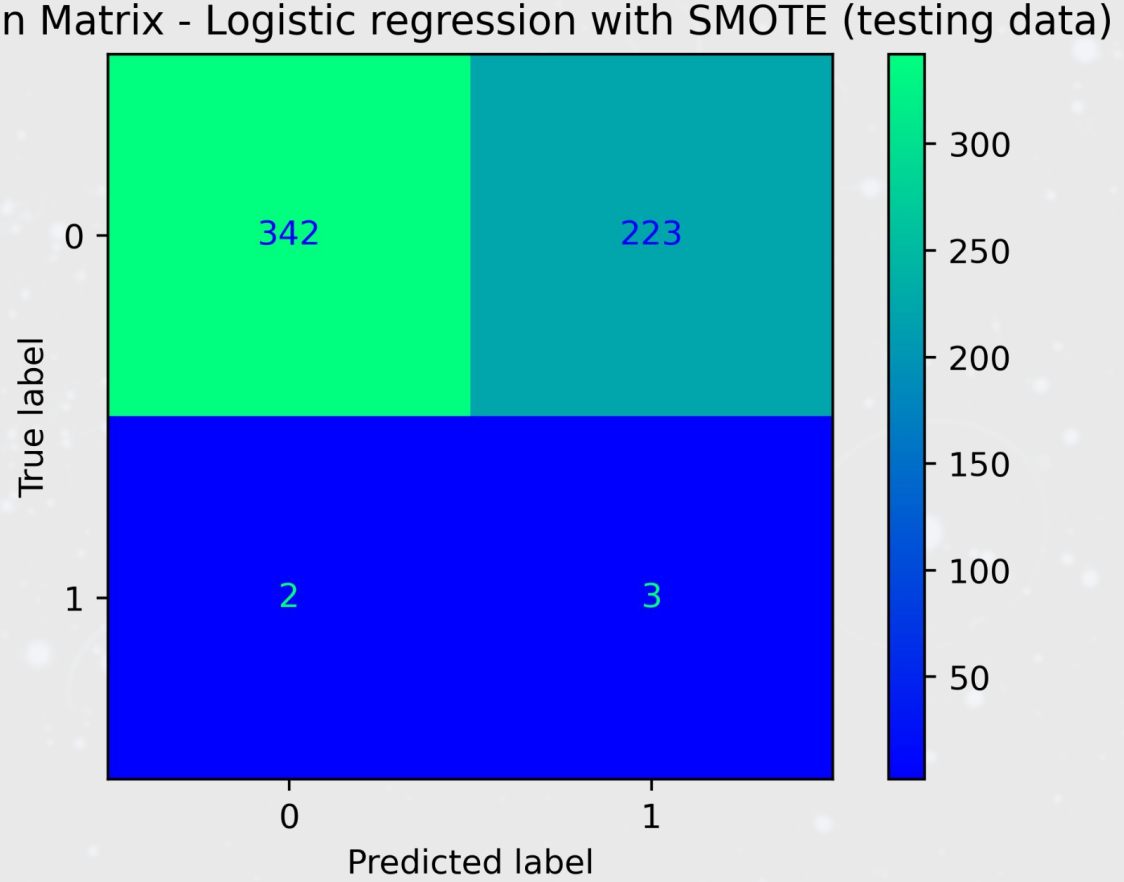
Precision: 1



MODELING - LOGISTIC REGRESSION

On test (trained on SMOTE)

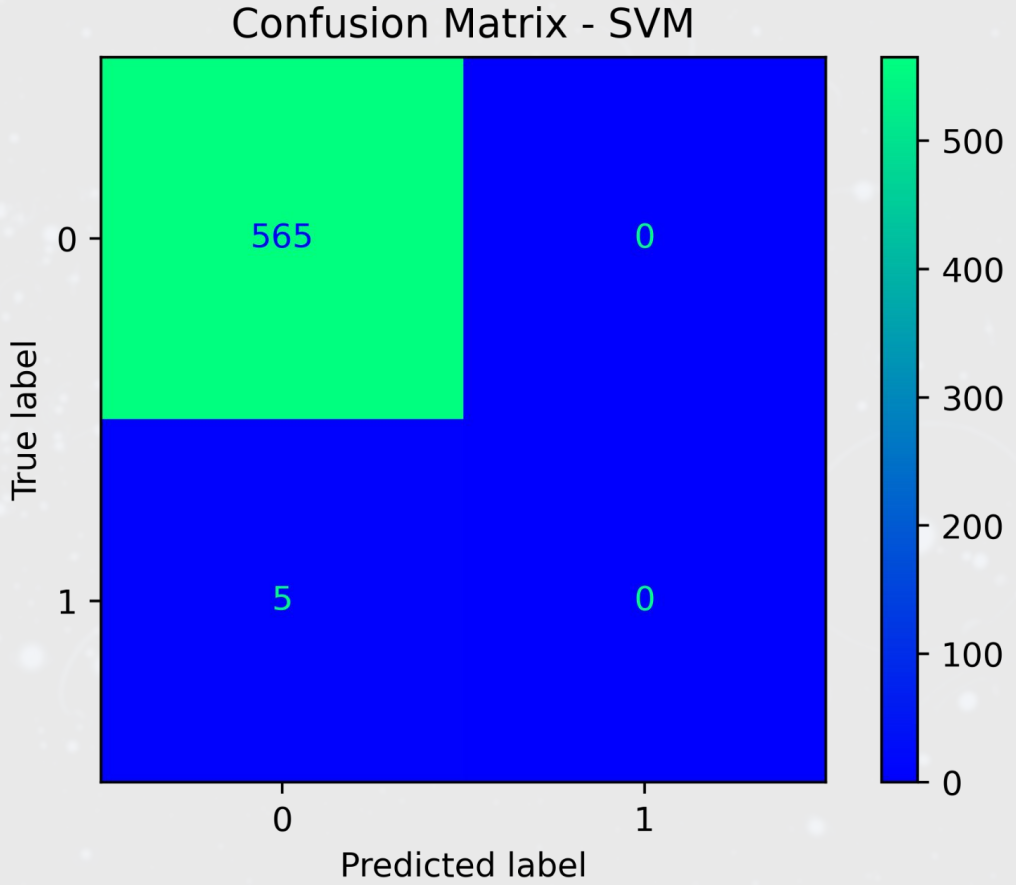
Precision: 0.6



MODELING - SUPPORT VECTOR MACHINE

Before SMOTE

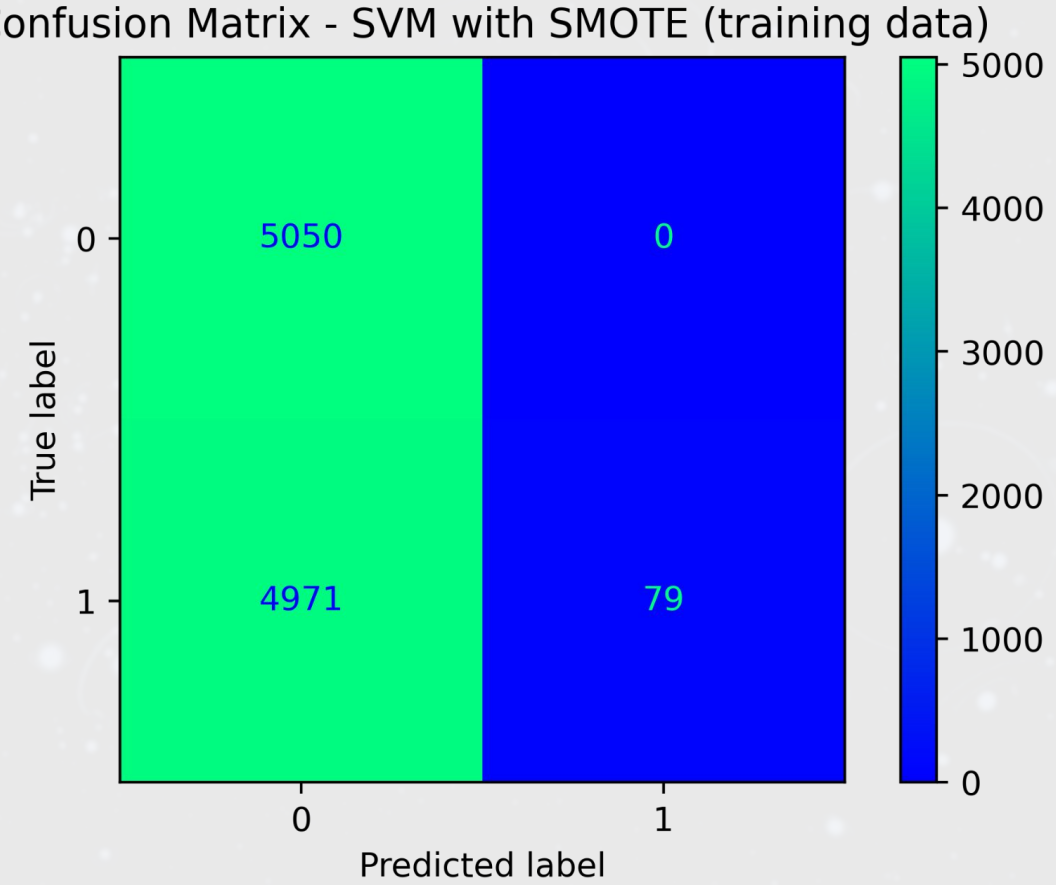
Precision: 0.0



MODELING - SUPPORT VECTOR MACHINE

On SMOTE data

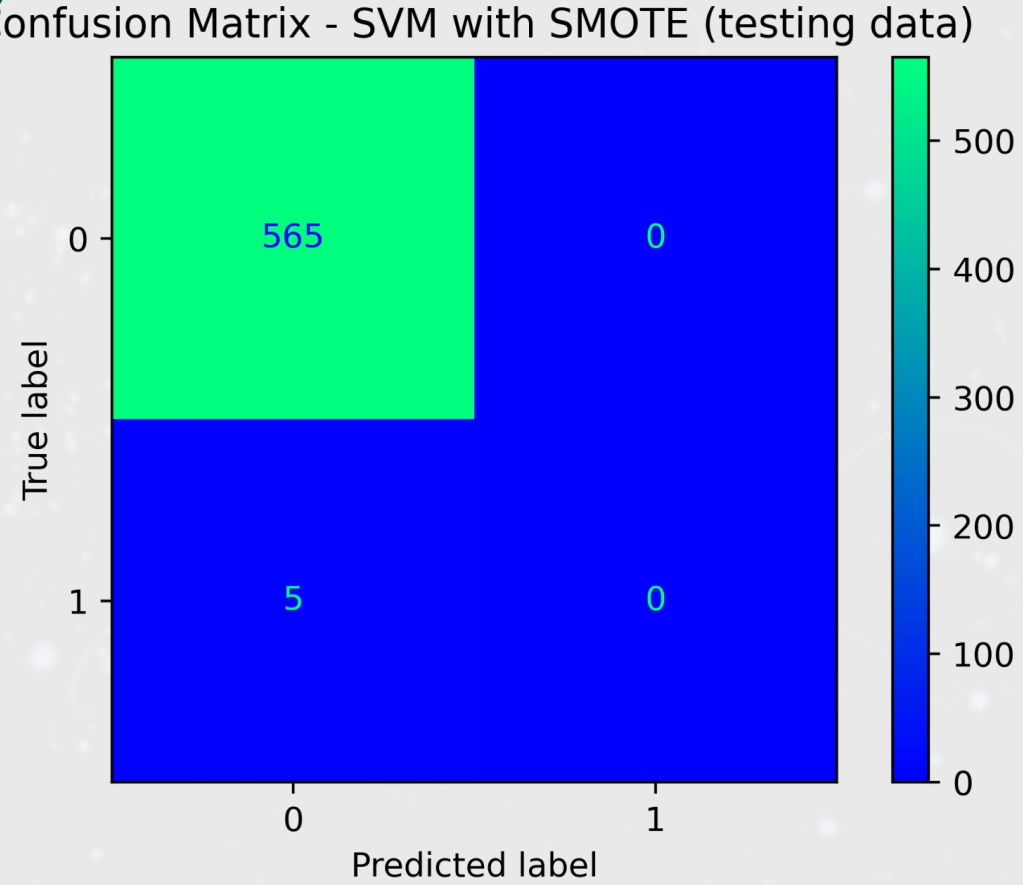
Precision: 0.02



MODELING - SUPPORT VECTOR MACHINE

On test (trained on SMOTE)

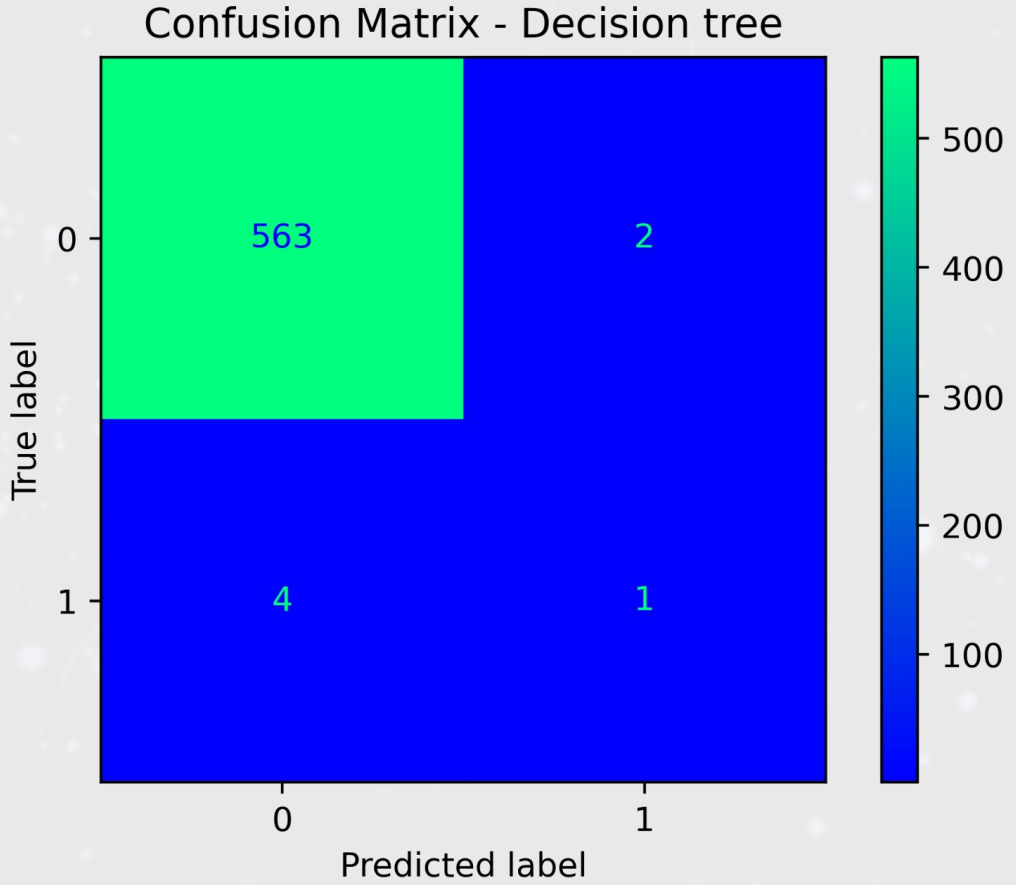
Precision: 0.0



MODELING - DECISION TREE

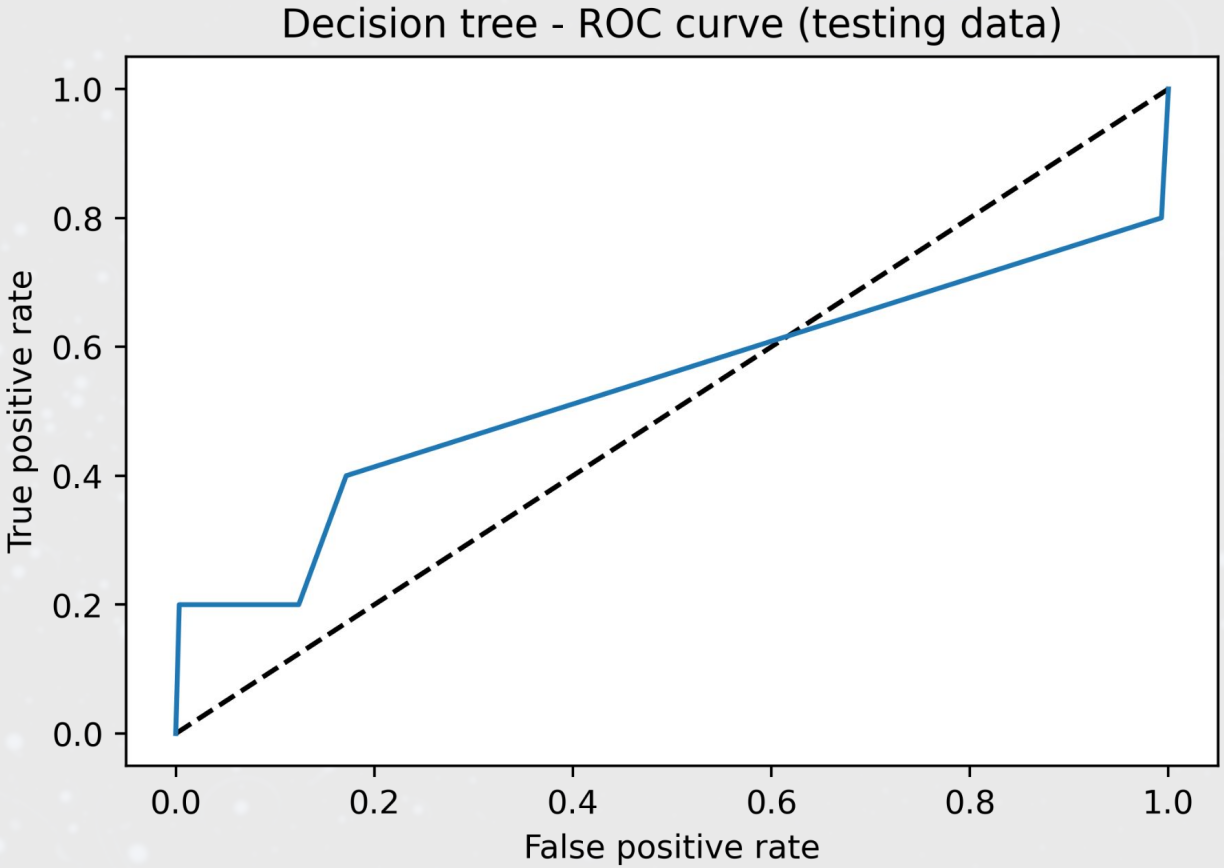
Before SMOTE

Precision: 0.2



MODELING - DECISION TREE

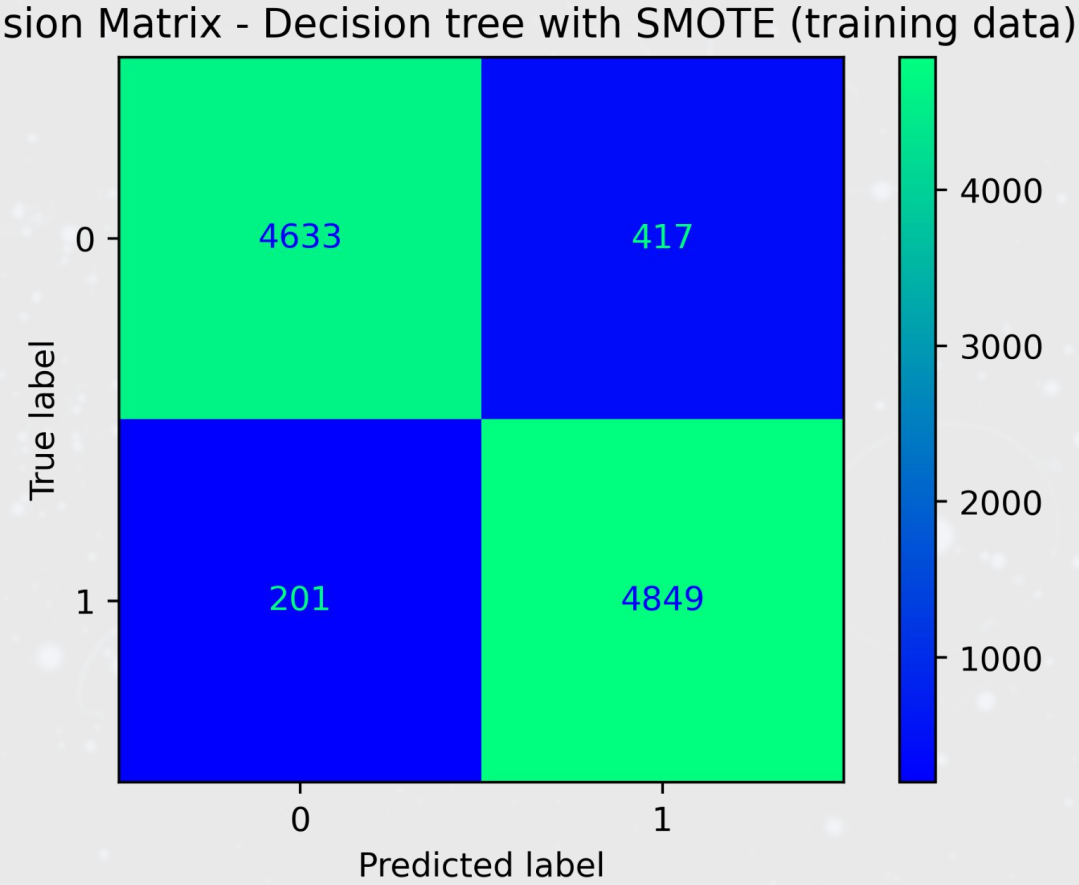
Before SMOTE - ROC curve



MODELING - DECISION TREE

On SMOTE data

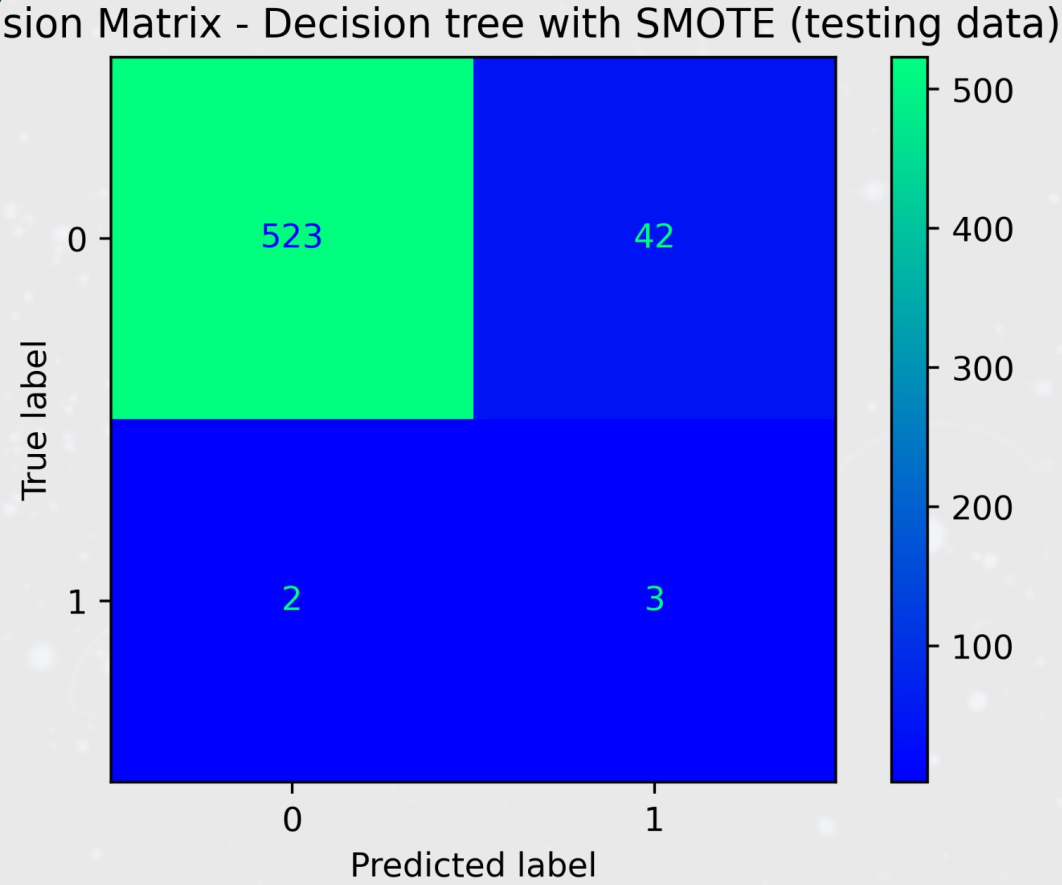
Precision: 0.96



MODELING - SUPPORT VECTOR MACHINE

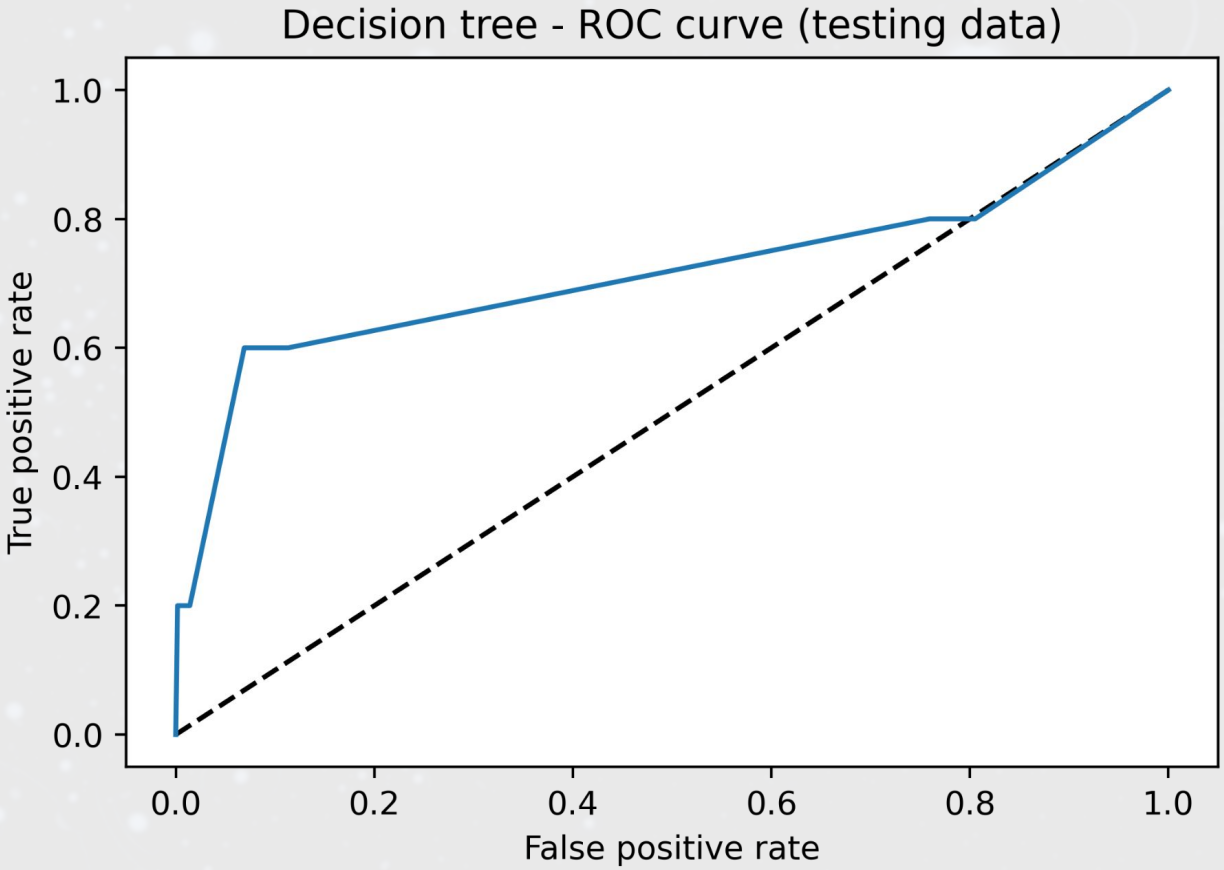
On test (trained on SMOTE)

Precision: 0.6



MODELING - DECISION TREE

Before SMOTE - ROC curve



CONCLUSION

1. ***Logistic regression***
2. *Decision tree*
3. *Support Vector Machine*

- *No model is actually satisfying*
- ***Testing** set with 5 observations might be **too small** as well*
- *Models performing well on SMOTE data still **misclassify almost half** of exoplanets*

NEXT STEPS

- *Get **more data** from the Kepler mission*
- *Try some **more**, different **models** (random forest)*
- *Try completely **different modeling techniques** (deep learning)*
- *Try **other techniques** for dealing with **imbalance***

*Hopefully we can try some of these in the **written report**.*