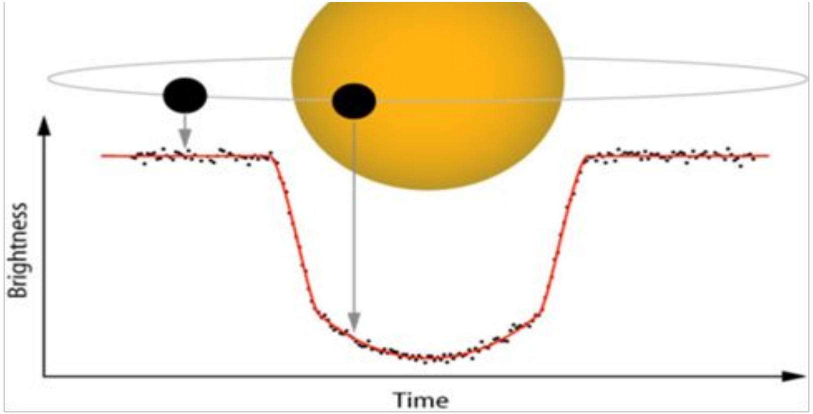
Searching For Exoplanets

Introduction: What are Exoplanets?

- Exoplanets are planets orbiting a star other than our Sun.
- Exoplanets were first discovered in 1992.
- They vary in size and orbital period length.
- In this project we aimed to use machine learning and AI to search for exoplanets using transit photometry.

Introduction: What are Exoplanets?

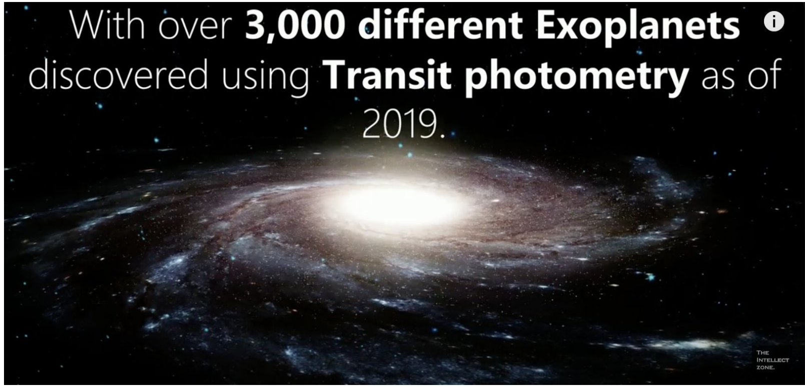
- Transit photometry: indirectly detecting exoplanets based on their effect on star brightness
- Exoplanet detection and confirmation can lead to further observations of the object.



Task

Using the power of AI to detect exoplanets in space

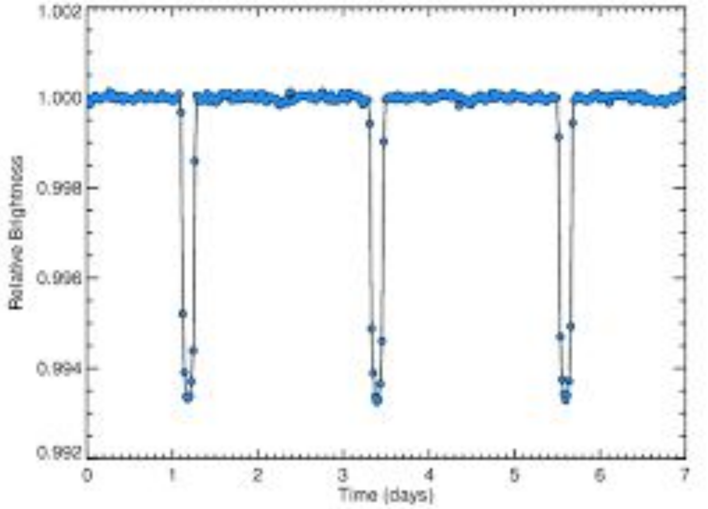
1. AI should be able to classify a star that has or does not have an exoplanet orbiting it
2. Using transit photometry data
3. Models used:
   a. KNN (k-nearest-model)
   b. Logistic Regression
   c. Decision Trees
   d. Convolution Neural Network (CNN)



With over **3,000 different Exoplanets** discovered using **Transit photometry** as of 2019.

Data

1. Training Data:
   a. Labeled data fed into AI which based the classifications
   b. Labeled graphs from NASA
2. Testing Data
   a. Unclassified data: AI would classify based from the training data
   b. Unlabeled graphs: Classified as a star with or without an exoplanet

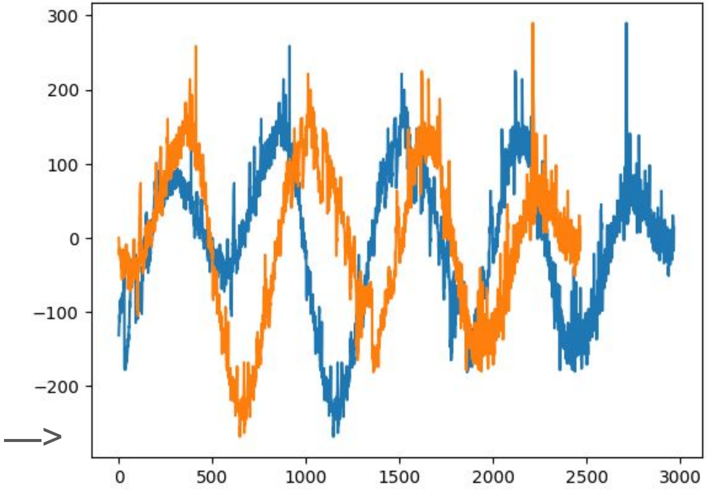


Manual Efforts: Folding

To manually determine if something is an exoplanet, we can use something called folding.

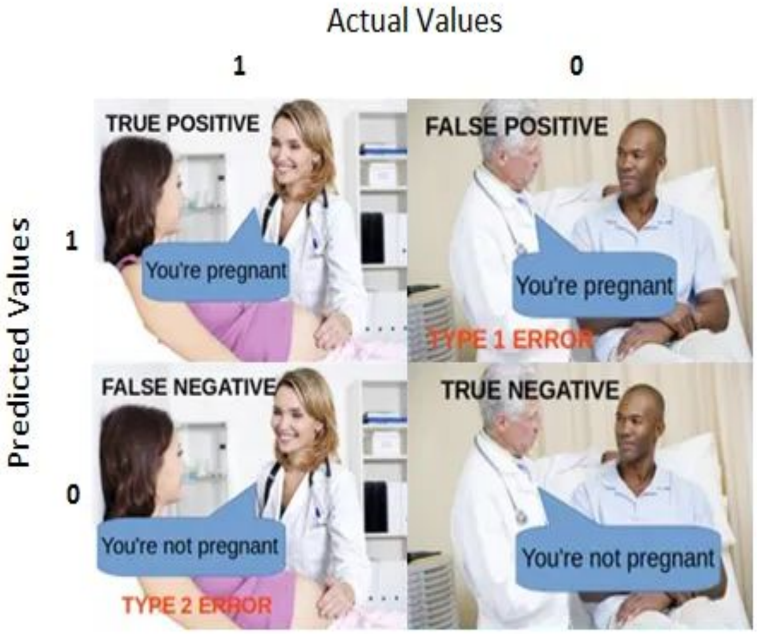
We do this by plotting all the periods on top of each other to see if there's a consistent trend.

We want to see if the dips occur roughly at the same places.



the two are overlaid —>

Confusion Matrix Explanation



- Used to access the accuracy of your model.
  - If the actual value = predicted, it is accurate.
  - If the actual value != predicted, it is inaccurate
- Using accuracy, precision, & recall
  - Accuracy = overall ‘correctness’ of the model
  - Precision = accuracy of positive predictions
  - Recall = ability of model to find all positive cases.
- Pros & Cons
  - Determine model’s accuracy
  - Not good with multiclass
  - Can be misleading

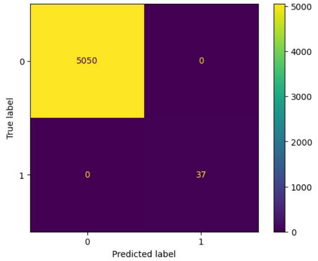
Sofiya Zuykova

First Attempt Using ML Models

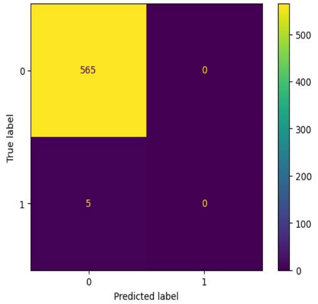
KNN, Logistic Regression, and Decision Trees

In order to identify which stars have exoplanets, we can try using different ML models.

KNN

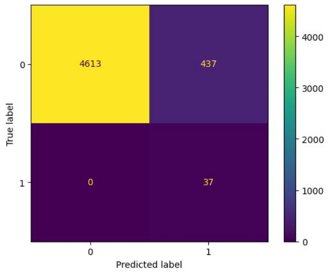


On the training data, the model has a 100% accuracy because this is what the model is trained on.

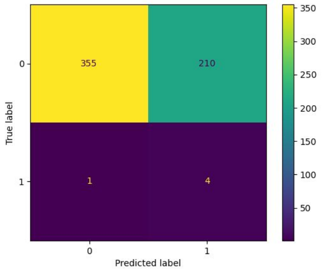


On the test data, it always predicts an exoplanet because the ratio of exoplanet to non-exoplanet is so high, and it can be 99% accurate if it just guesses non-exoplanet

Logistic Regression



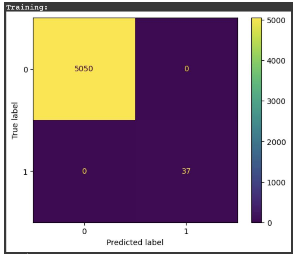
The Logistic Regression model specializes in classify different objects into categories, but the total accuracy of the training set is much lower than KNN



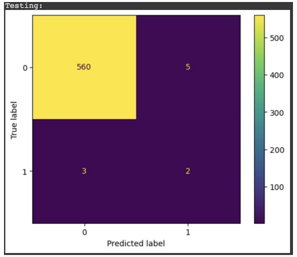
The test set accuracy is also much lower, but this model has higher accuracy when considering the ratio of false negatives to true positives, but also has a lot false positives

First Attempt Using ML Models

Decision Tree



The Decision Tree model trained its “decisions” on the training model and thus, has a 100% accuracy just like the KNN model



Unlike the KNN and Logistic Regression model, the Decision Tree doesn’t just guess exoplanet and it doesn’t such high amounts of error though it still has a higher amount of false positives compared to true positives

Although these ML models normally work, the dataset we have is heavily biased with a high amount of non-exoplanets. To get better results we must modify our dataset

Data Augmentation - Normalization

Because our range of flux values vary very wildly, we can use normalization to make our data more concise and easier.

In this case, we scaled our data so that all the points would lie between 0 and 1.

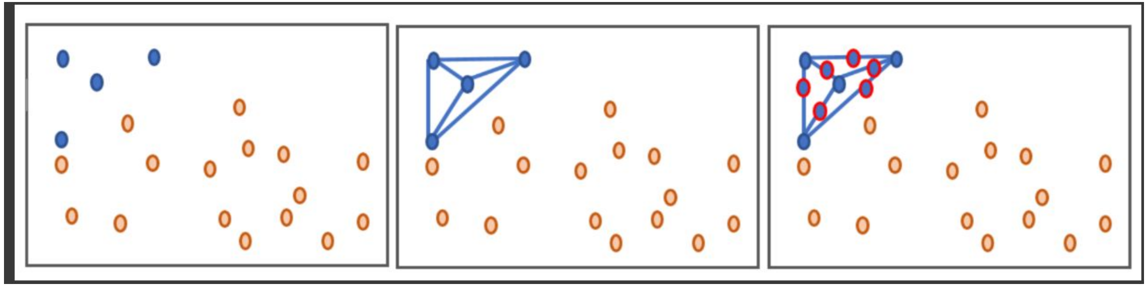
For example, if we were to normalize this list of values…

[0, 1, 2, 3, 4, 5]

The normalized list would be:

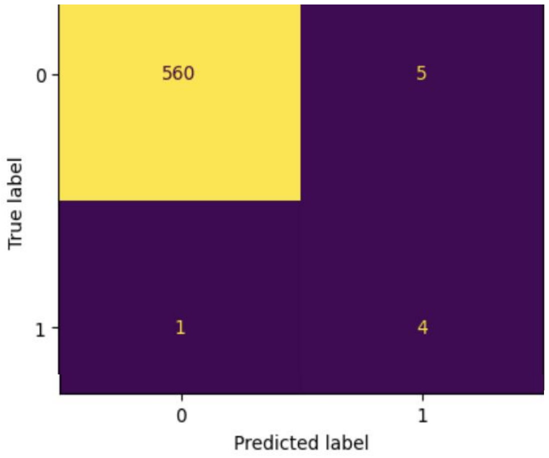
[0, 0.2, 0.4, 0.6, 0.8, 1]

Data Augmentation - SMOTE



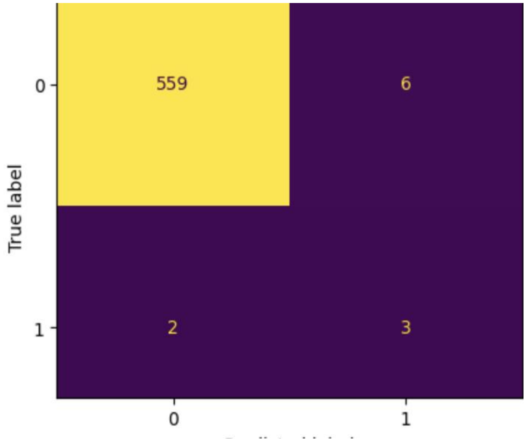
The blue points are confirmed exoplanets in our hypothetical scenario. What SMOTE does, is that it connects the dots and places new synthetic data points at about ⅔ of the distance from point to point. This new addition assists the AI to help us as much as possible to help us make more observations.

Second Attempt Using ML Models (With Data Augmentation)



Logistic Regression

- Correctly predicted 99% of the non-exoplanets, which is better than the non-augmented model's 62%
- It correctly identified 4 of the 5 exoplanets, which is the same as the model without augmentation



Decision Trees

- This model with augmented data correctly identified the same amount of non-exoplanets as the one without augmentation
- It identified 3 of the 5 exoplanets, which is better than the non-augmented model's 1 out of 5

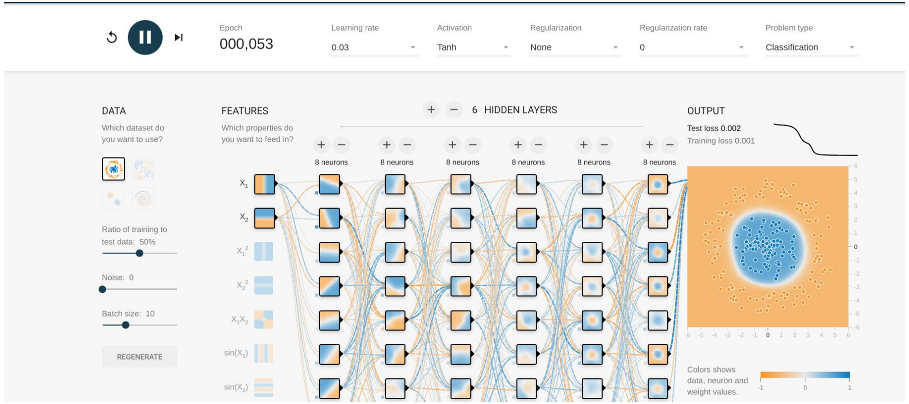
Neural Networks and CNNs

What are Neural Networks and CNNs (Convolutional neural networks)?

A neural network is a replication of the human brain that uses nodes as neurons. They come to a conclusion using logic rather than a pre written program. Convolutional neural networks are networks that specialize in image recognition and classifications.

Where are these two useful?

Neural Networks	Convolutional Neural Networks
• Medical diagnosis • Targeted marketing • Financial predictions • Self-driving cars • Chat AI (Chat GPT, Gemini, website landing page helpers, etc)	• Facial recognition • Image classification • Object recognition • Recommender systems • Health risk assessment • Autonomous systems



Epoch 000,053

Learning rate 0.03

Activation Tanh

Regularization None

Regularization rate 0

Problem type Classification

DATA

Which dataset do you want to use?

Ratio of training to test data: 50%

Noise: 0

Batch size: 10

FEATURES

Which properties do you want to feed in?

6 HIDDEN LAYERS

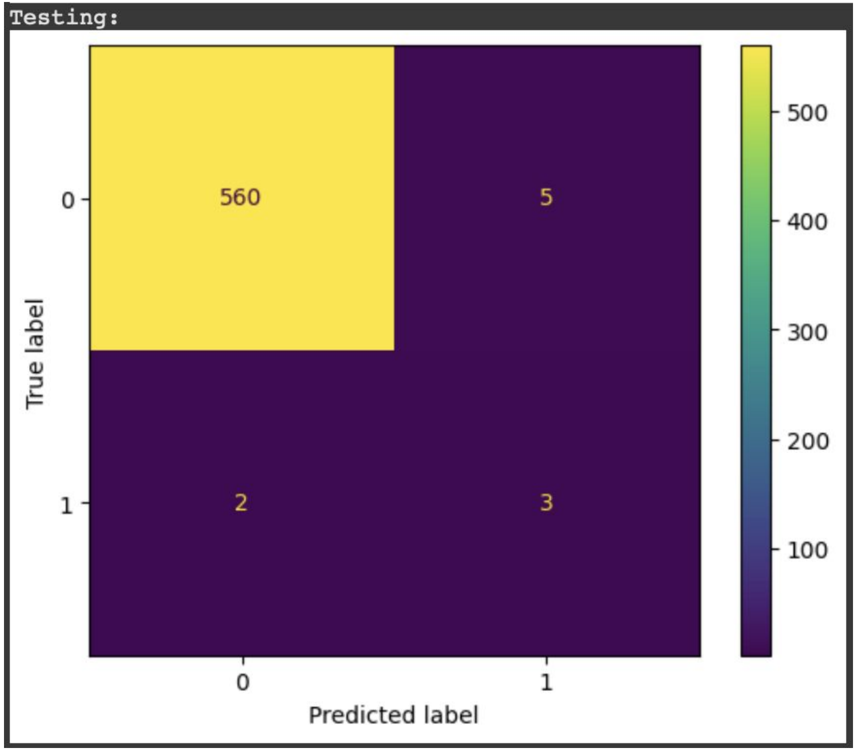
OUTPUT

Test loss 0.002

Training loss 0.001

Colors shows data, neuron and weight values.

How it applied to our project (NN and CNN continued)



```
Testing:
```

As we can see, the AI used the flux data to predict the classification type of the data points.

Conclusion

- In this project, we have learned how to train and test various AI models to detect exoplanets using transit photometry data.
- Models were assessed using a confusion matrix.
- Accuracy varied among methods.
- As we increase the complexity of our models(normalization, augmentation, hidden layers in CNNs), the prediction accuracy increases.