INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD

C1 REPORT

ON

“PREDICTING HABITABLE EXOPLANETS IN DIFFERENT STAR-SYSTEM USING DEEP LEARNING”

SUBMITTED BY
YASH PATEL (MIT2021090).

UNDER THE SUPERVISION OF
DR. SONALI AGARWAL

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD

(A UNIVERSITY ESTABLISHED UNDER SEC.3 OF UGC ACT, 1956 VIDE NOTIFICATION NO. F.9-4/99-U.3 DATED 04.08.2000 OF THE GOVT. OF INDIA)
A CENTRE OF EXCELLENCE IN INFORMATION TECHNOLOGY ESTABLISHED BY GOVT.OF INDIA

APRIL, 2022

CANDIDATE’S DECLARATION

I hereby declare that the work presented in this study report entitled **“PREDICTING HABITABLE EXOPLANETS IN DIFFERENT STAR SYSTEMS USING DEEP LEARNING”**, submitted towards fulfillment of C2 Exam (part of project) of MTech (IT) at Indian Institute of Information Technology, Allahabad, is an authenticated record of original work carried out from January 2022 to May 2022 under the guidance of **Dr. Sonali Agarwal**. Due acknowledgements have been made in the text to all other resources used. The study was done in full compliance with the requirements and constraints of the prescribed curriculum.

Place: Allahabad	YASH PATEL
Date: 10/05/2022	MIT2021090

CERTIFICATE

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date:
Place: Allahabad

Dr. Sonali Agarwal
Associate Professor, IIITA

Acknowledgement

I have tried my best to present this state of the art report in a complete manner without failing the deadlines. However, it would not have been possible without the kind support and help of many individuals. We would like to extend our sincere thanks to all of them.

I am very highly indebted to **Dr. Sonali Agarwal** for their guidance and constant supervision as well as for providing necessary information concerning to implementation. I am very thankful to them for their support in this comprehensive in-depth study.

Motivation

Mankind has been looking up at the stars for centuries, wondering what lies out there in deep space, and if there are other civilizations like ours that exists.

Observing exoplanets is no easy task. In order to give a complete picture of an exoplanets, observational data from multiple missions is needed to be combined.

For instance, the method of radial velocity gives minimum mass of the exoplanet and the distance of planet from its host/parent star, transit photometry can provide us with the radius of the planet and gravitational lensing can give us the information like mass.

Therefore, handling this large amount of data collected by different missions and manually handling it can be a tedious work which would require a lot of manpower, money and time. However, all those setbacks can be overcome by using machine learning and deep learning to process large amount of raw data collected by these missions and building a model that will feed on that data and provide us with the desired result without much utilization of resources.

We can furthermore help to develop an interface which will help researchers/scientists all over the globe to easily feed the desired features into the well-designed user interface to get the desired output as new data is being collected by these missions. Which will make it easier for researchers with no or less knowledge about these models to easily make use of this domain thus saving resources.

Objective

Objective of this project work is to make use of concept of Artificial Neural Networks from deep learning to solve the problem of prediction of habitability index of exoplanets which is not effective manually with the exponential increase in the astronomical data collected by multiple missions.

Our focus will be on achieving the better efficiency in our prediction model as compared to previously build models in this field.

Moreover, there is only handful of work done in this field using deep learning models so it's a quite new area of research to apply the concept of deep learning (Neural Networks).

With so many missions now completed and tones of data collected by them and many more future missions like James Webb Space Telescope and PLAnetary Transits and Oscillations of stars (*PLATO*) in the line there will be exponential increase in astronomical data collected by them. Therefore, this is the right time to explore deep learning which requires huge amount of data in the field of astronomy.

Table of Contents

Abstract

Since the beginning of cosmological expansion, the observable universe is estimated to be about the diameter of 93 billion light years. In this vast cosmic space, there might be a higher probability of the existence of earth size celestial bodies or similar habitable zones. This popular interest led to embracing the vision to explore our Milky Way galaxy. In 2009, NASA launched the Kepler Space Telescope which studied a small patch of area which consists of about half-million stars. Our proposed way to detect habitability consists of statistical approach by using Deep-learning algorithms on Kepler and TESS dataset. Our prime centre is to build the increased set of parameters over which habitability of an exoplanet depends and furthermore utilizing a dataset with an increased set of the star systems which will additionally bring about expanded precision and accuracy of the prediction.

Introduction

The observable universe is around 46.5 billion light years in radius and contains around 200 billion to 2 trillion cosmic systems. In this inconceivability of room here's our home called The Milky Way. Cosmologists assessed that there are about 250±150 billion stars in the Milky Way alone and stars like the Sun make up approximately 10 percent of all-stars for example around 20±15 billion stars. A rocket was launched by NASA on March 7, 2009, which was a part of NASA's Discovery Program by a telescope named Kepler Space Telescope to find Earth-size planets circling different star systems.

In the mid-90s, the researcher’s began finding planets around other habitable planets. As of late distributed paper "Planetary competitors saw by Kepler VIII" recognized a few stars, called exoplanets, utilizing Doppler spectroscopy, at times called the spiral speed technique, and ordinarily known as the wobble strategy. As of April 2016, 582 exoplanets (about 29.6% of the all known at that point) were found utilizing this strategy.

Objective of this project work is to make use of concept of Artificial Neural Networks from deep learning to solve the problem of prediction of habitability index of exoplanets which is not effective manually with the exponential increase in the astronomical data collected by multiple missions.

Our focus will be on achieving the better efficiency in our prediction model as compared to previously build models in this field.

Related Work:

Research Work	Methodology	Approach	Limitations	Future Work
Classifying Exoplanets as Potentially Habitable Using Machine Learning. [Karan Hora]	Six supervised learning algorithms for the classification: 2 Decision Tree, CART and random forest, Logistic regression, Feed-Forward Neural Network, SVM and Naïve Bayes.	Used Habitable Exoplanets Catalog contains 1,943 exoplanets out of which only 31 are categorized as potentially habitable. Dividing the data in the ratio of 70:30 for training and testing.	Only handful of exoplanet data containing only 1943 datapoints. Large gap in the ratio of potential habitable and non-habitable exoplanets i.e., 1943:31 only.	Models can be further optimized as the available data from various missions increases in size.
Finding Habitable Exo Planets Using Boosting Algorithm. [Md. Mashfiq Rahman]	Used 2 classification-based machine learning algorithm i.e., KNN classifier and Decision tree classifier. Also boosting algorithms are used to increase efficiency of above algorithms namely Ada boosting and XGA (extreme gradient boosting) algorithm.	Dataset of 3,874 known exoplanets is used. Which is divided in 80:20 ration for training and testing. 80% of training data is further divided into 70:30 ratio for validation out of which 70% is used for creating model and remaining 30 to fit the model.	Models are built only on limited i.e., 5 parameters: Habitability Zone, period, mass and radius, metallicity and eccentricity.	As size of dataset increases efficiency of models will also increase. And set an unsupervised program to collect data continuously and analyze it.
Predicting habitable exoplanets from NASA's Kepler mission data using Machine Learning.	Used SVM model along with forward feature selection and reverse feature elimination algorithms.	Initially 2,373 planetary data was used in the ratio of 50:20:30 for training, dev and testing. And later model was deployed on	14 stellar and planetary parameters were used in the model out of 140. Although not all of them affect the result but other	With more upcoming space telescope missions as the available data increases so will the efficiency of

[Rajeev Misra]		whole dataset of 9,564 planets for prediction.	parameters such as few optical parameters could have been considered.	machine learning models.
CEESA meets machine learning: A Constant Elasticity Earth Similarity Approach to habitability and classification of exoplanets. [S. Basak, S. Mathur]	Developed new model namely CEESA which uses parameters radius, density, surface temperature, escape velocity and eccentricity to calculate the habitability score of exoplanets.	This model overcame the drawback of '0' values in the dataset by using models like KNN to further optimize the existing models like CD-HPF. ANN and fuzzy NN were also experimented.	## No information about dataset provided.	With increase in available data efficacy of the model will also increase.
Identifying Exoplanets Using Deep Learning and Predicting Their Likelihood of Habitability. [Somil Mathur]	CNN is implemented on raw data of transit signals of extrasolar objects to classify them as exoplanets or non-exoplanets and the various machine learning models were implemented to predict habitability.	CNN is implemented in three stages: without preprocessing, Savitsky Golay filter and Gaussian filter. Data was divided in the ratio of 80:20. And then KNN, SVM and Random Forest was implemented to categorize planets as Mesoplanet, Psychroplanet and non-habitable.	## No information about dataset provided.	Further classification of outcome other than Mesoplanet, Psychroplanet and non-habitable can give better insight into the habitability index of exoplanets.
Habitability classification of exoplanets: a machine learning insight.	Investigate the existing machine learning algorithms, such as Naive Bayes,	Machine Learning algorithm are investigated on the dataset containing 3,800	Limited dataset of 3,800 objects from PHL-EC catalogue.	As available dataset of these exoplanets increases with more upcoming

[S Basak, A Mathur]	Linear Discriminant Analysis, Support Vector Machine, KNN, Decision Tree, Random Forest and XGBoost. Exploration of neural network models like RWNN, GAN and fusion net.	data points of extrasolar objects. And then neural network models are used with activation function used being Leaky RELU.		future missions' efficiency and working of other algorithms can be tested.
Evolution of novel activation functions in neural network training for astronomy data: habitability classification of exoplanets. **[Snehanshu Saha]**	Instead of ANN RWNN neural network algorithm is used. And the activation function being used is SBAF (Saha-Bora Activation function) along with RELU, Sigmoid and A-RELU being used for approximating these functions.	PHL-EC data catalogue of 3,800 objects is being used on over 68 features of which 13 are categorical and remaining continuous. For sigmoid AF 20 hidden neurons are used while for SBAF, RELU and A-RELU 11 hidden neurons were used and tested over different subsets of features.	Number of exoplanets dataset has considerably increased since publication.	As the number of observed exoplanets and their biosignature features increases with upcoming missions like JWST and PLATO the habitability index prediction integrity will increase.
Habitability of exoplanets using DEEP LEARNING. **[Rutuja Jagtap]**	Uses ASTRONET architecture model for classification of exoplanets as habitable or not. Deep convolutional network is used along with sigmoid and ReLU	Data collected by NASA's Kepler mission and TESS is being used. First raw data is fed into model to predict if the object is truly an exoplanet or the result of an anomaly.	## No information about dataset provided.	With upcoming data, which can be directly fed into the system to evaluate the model prediction.

	activation function.	Then this data is fed into model to predict habitability of the exoplanet.		
Exoplanet Hunting in Deep Space with Machine Learning. [Shivam P Singh]	The dataset used in this project was obtained from National Aeronautics and Space Administration, "Kepler and K2”. Machine learning algorithms used for prediction are KNN, SVM, Linear Discriminant Analysis (LDA), Naïve Bayes, Decision Tree and Random Forest.	Data collected by NASA’s Kepler mission is being used. Data preprocessing and Normalization is used and then SMOTE technique Was used to overcome class imbalance. And then ML classification models were implemented.	## No information about dataset provided.	Exoplanet’s data would be present in large number in future. Machine learning would assist us in gathering the existing knowledge about habitability via machine learning model and using it on new data to make the list of habitable planets. More accurate model would be possible when more training data about habitable planets become available.
A convolutional neural network (CNN) based ensemble model for exoplanet detection. [Ishaani Priyadarshini]	Used Ensemble-CNN for exoplanet detection. And then compared the result on accuracy, precision, sensitivity and specificity with other machine learning models like Logistic Regression, SVM, Decision Tree, Multilayer Perceptron, Random Forest and CNN.	NASA’s data collected by different missions comprises of 5,087 data points for training and 570 data points for testing on 3,198 features. Features from COL 1 to COL 3197 consist of flux values of each star system over which Ensemble-CNN is applied.	Dataset 0f about 5,157-star system is used with 3,198 features ranging from FLUX1 to FLUX3197. Out of 5,157 data points 5,087 are used for training and remaining 570 for testing.	It is not easy to detect light reflected from a planet’s atmosphere. It would be interesting to explore other features and combine them with Artificial Intelligence techniques for exoplanet detection.

Challenges and Research Gaps:

Deep learning concept is quite a new field to explore in the arena of astronomy specially when it comes to this particular field of predicting habitability index of exoplanets. There’s a limited work done using neural networks primarily due to insufficient data availability.

But in the recent past with completion of NASA’s Keppler mission and the launch of TESS mission there’s been enough data accumulated about extrasolar planets to explore the concepts of deep neural networks in this field.

From this limited exploration of deep neural networks only basic concepts are being applied for example till now only Saha-Bora (SBAF), leaky-RELU and A-RELU activation functions are being used in backpropagation and updating of weights of networks in the models. There are so many advanced activation functions like SoftMax, Swish and Softplus present which are yet to be explored.

There’s also a problem associated with these already used activation functions like SABF and leaky-RELU. Since SBAF is an expansion of Sigmoid activation function only and the problem associated with sigmoid function is that of vanishing gradient however, that drawback has been overcome in SBAF.

But still, there’s a drawback that can affect the efficiency of the model which is that SBAF function is not adaptive in nature which means that the value of activation function will remain the same for every epoch and every layer in a neural network.

Same is the issue associated with leaky-RELU activation function if the value of constant in this function is set low then it will result in a vanishing gradient error and efficiency of the algorithm can be seriously affected.

Problem Formulation:

As our Kepler Mission itself collected a huge amount of exoplanet’s data and this volume of data will keep on increasing with future exploration missions e.g., James Webb Space Telescope (already launched in December 2021) and PLATO (scheduled for launch in 2024). Therefore, in near future there will be explosion of data collected from these space missions and hence implementing statistical analysis on such huge datasets requires complex computational manpower. So, to tackle this problem, we came up with mixed approach of astronomy and deep learning.

Therefore, we will be implementing deep learning algorithms (Neural Networks) on the dataset collected by Keppler mission and TESS mission to predict the habitability index of exoplanets. In this work we will explore the advance novel activation functions along with the integration of several other ideas like optimizers and hyperparameter optimizers leading to the implementation and classification of exoplanets as habitable of non-habitable.

Methodology:

In the mid-90s, the researcher’s began finding planets around other habitable planets. As of late distributed paper "Planetary competitors saw by Kepler VIII" recognized a few stars, called exoplanets, utilizing Doppler spectroscopy, at times called the spiral speed technique, and ordinarily known as the wobble strategy. As of April 2016, 582 exoplanets (about 29.6% of the all-out known at that point) were found utilizing this strategy.

Later on, different procedures were presented, for example, ·

- **Direct Imaging:** Direct imaging is an extremely troublesome and restricting strategy for finding exoplanets. Above all else, the star framework must be moderately near Earth. Next, the exoplanets in that framework must be far enough from the star with the goal that astronomers can recognize them from the star's glare. ·
- **Transit Method:** The travel technique depends on the perception of a star's little drop in brilliance, that happens when the circle (ran line) of one of the star's planets passes ('travels') before the star The measure of light lost- ordinarily somewhere in the range of 0.01% and 1%-relies upon the extents of the star and the planet and the duration of revolution. ·
- **Micro lensing:** This technique depends on the gravitational power of far-off objects to twist and concentrate light originating from a star. As a planet goes before the star comparative with the onlooker (for example makes travel), the light plunges quantifiably, which would then be able to be utilized to decide the nearness of a planet.

The exoplanets and their features gathered from all these techniques is collectively used to prepare a dataset with all objects and features which will be fed into the deep neural network model.
Our proposed model will explore some of the advanced and unexplored concepts of deep learning like advance activation functions like SoftMax, Swish and Softplus and various optimizers like ADAM and ADAGRAD with different number of hidden layers and epochs.

Requirements:

Dataset Description:

As of 25th February 2022, the current tally to total discovered exoplanet rose to 4,935.
This huge dataset of discovered extrasolar planets further have a large feature set of roughly 68 parameters.

These features can be broadly divided into two categories namely:

1. Categorical
2. Continuous

Both of these can contain two subsets of parameters namely:

1. Planetary features: Such as planet radius, temperature, mass, time period of rotation, time period of revolution, etc.
2. Stellar features: Such as star radius, mass, temperature, etc.

Hardware requirements:

Hardware requirements can comprise of high-end features like:

1. High performance processors
2. High RAM requirement probably greater than 64GB
3. HDD with SDD support
4. Graphical Processing Units (GPU’s) like NVidia TitanXPascal
5. Cooling system to maintain temperature of the system in control

Software requirements:

Software requirements may include:

1. Choosing the right Operating System: Generally, Ubuntu OS is preferred for these types of projects on Deep Learning.
2. Installing programming language: R or Python programming language is preferrable.
3. Installing other packages: Various deep learning tools can be set up.

Implementation Plan:

Rough idea on implementation timeline for this proposed project on predicting habitable exoplanets using deep learning:

Time Period	Description
1st Feb to 28th Feb	Project selection, Research work, literature survey, Report making.
1st March to 15th April	Data pre-processing, data extraction, hyperparameter optimization, parameter selection.
15th April to …	Model building, training and testing, Outcomes.

Expected Outcomes:

Outcome of our proposed model will be based on four different parameters:

1. **Accuracy:** Accuracy is a metric that may be used for evaluating classification models. It emphasizes how often an algorithm classifies data correctly. It may be defined as the number of correctly predicted data points with respect to the total number of data points. Given a confusion matrix, accuracy may be defined as the sum of True Negative and True Positives combined over the sum of True Negative (TN), True Positive (TP), False Negative (FN), and False Positive (FP).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2. **Precision:** Precision is a metric that is popularly used in classification, information retrieval, and pattern recognition. It may be defined as the number of relevant observations with respect to retrieved observations. Given a confusion matrix, Precision may be calculated by True Positives with respect to the total number of True Positives and False Negatives combined.

$$\text{Precision} = \frac{TP}{TP + FP}$$

3. **Sensitivity:** Sensitivity metric is defined as the ratio of actual positive events that got predicted as positive. It is sometimes also referred to as recall. Given a confusion matrix, sensitivity may be calculated by True positive value with respect to the sum of True Positive and False Negative combined)

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

4. **Specificity:** Specificity metric is defined as the ratio of actual negatives that got predicted as negative. Given a confusion matrix, specificity would be calculated by True Negatives with respect to the sum of True Negatives and False Positives combined.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Data Pre-processing:

1. **Exploratory data analysis (EDA):**
   The PHL-EC dataset consist of 3875 planetary catalogues from different star systems with over 69 different parameters. These parameters are broadly divided into two categories having 56 features as numerical and 13 categorical features.
   This section will deal with:
   - **Missing values**
   - **Distribution of numerical variables**
   - **Cardinality of categorical variables**
   - **Outliers**
   - **Relationship between dependent and independent features**

a. **Missing Values:**

A feature consisting of NULL values greater than 50% of the datapoints then those columns are needed to be removed as it will negatively affect our model.

However, feature set having smaller datapoints as NULL are needed to be handled carefully either by replacing them with mean or average of the column feature.

Following is the distribution of missing or NAN values from all of the available feature sets:

```
RangeIndex: 3875 entries, 0 to 3874
Data columns (total 69 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   P. Name                    3875 non-null   object
 1   P. Name Kepler             2328 non-null   object
 2   P. Name KOI                933 non-null    float64
 3   P. Zone Class              3829 non-null   object
 4   P. Mass Class              3869 non-null   object
 5   P. Composition Class       3834 non-null   object
 6   P. Atmosphere Class        3790 non-null   object
 7   P. Habitable Class         3875 non-null   object
 8   P. Min Mass (EU)           1148 non-null   float64
 9   P. Mass (EU)               3842 non-null   float64
 10  P. Max Mass (EU)           0 non-null      float64
 11  P. Radius (EU)             3863 non-null   float64
 12  P. Density (EU)            3834 non-null   float64
 13  P. Gravity (EU)            3834 non-null   float64
 14  P. Esc Vel (EU)            3834 non-null   float64
 15  P. SFlux Min (EU)          3875 non-null   object
 16  P. SFlux Mean (EU)         3875 non-null   object
 17  P. SFlux Max (EU)          3875 non-null   object
 18  P. Teq Min (K)             3829 non-null   float64
 19  P. Teq Mean (K)            3829 non-null   float64
 20  P. Teq Max (K)             3829 non-null   float64
 21  P. Ts Min (K)              1749 non-null   float64
 22  P. Ts Mean (K)             1749 non-null   float64
 23  P. Ts Max (K)              1749 non-null   float64
 24  P. Surf Press (EU)         3834 non-null   float64
 25  P. Mag                     3819 non-null   float64
 26  P. Appar Size (deg)        3863 non-null   float64
 27  P. Period (days)           3725 non-null   float64
 28  P. Sem Major Axis (AU)     3840 non-null   float64
 29  P. Eccentricity            3875 non-null   float64
 30  P. Mean Distance (AU)      3840 non-null   float64
 31  P. Inclination (deg)       724 non-null    float64
 32  P. Omega (deg)             3875 non-null   float64
 33  S. Name                    3875 non-null   object
 34  S. Name HD                 481 non-null    object
 35  S. Name HIP                521 non-null    object
 36  S. Constellation           3875 non-null   object
 37  S. Type                    3791 non-null   object
 38  S. Mass (SU)               3828 non-null   float64
 39  S. Radius (SU)             3763 non-null   float64
 40  S. Teff (K)                3776 non-null   float64
 41  S. Luminosity (SU)         3847 non-null   float64
 42  S. [Fe/H]                  2137 non-null   float64
 43  S. Age (Gyrs)              2322 non-null   float64
 44  S. Appar Mag               2946 non-null   float64
 45  S. Distance (pc)           2688 non-null   float64
 46  S. RA (hrs)                3875 non-null   float64
 47  S. DEC (deg)               3875 non-null   float64
 48  S. Mag from Planet         3829 non-null   float64
 49  S. Size from Planet (deg)  3747 non-null   float64
 50  S. No. Planets             3875 non-null   int64
 51  S. No. Planets HZ          3875 non-null   int64
 52  S. Hab Zone Min (AU)       3760 non-null   float64
 53  S. Hab Zone Max (AU)       3760 non-null   float64
 54  P. HZD                     3829 non-null   float64
 55  P. HZC                     3834 non-null   float64
 56  P. HZA                     3790 non-null   float64
 57  P. HZI                     3790 non-null   float64
 58  P. SPH                     1801 non-null   float64
 59  P. Int ESI                 3875 non-null   int64
 60  P. Surf ESI                3875 non-null   int64
 61  P. ESI                     3827 non-null   float64
 62  S. HabCat                  3875 non-null   int64
 63  P. Habitable               3875 non-null   int64
 64  P. Hab Moon                3875 non-null   int64
 65  P. Confirmed               3875 non-null   int64
 66  P. Disc. Method            3875 non-null   object
 67  P. Disc. Year              3875 non-null   object
 68  Unnamed: 68                2 non-null      float64
```

b. Distribution of numerical variables:

Distribution of numerical missing values in percentage was found to be:

```
P. Name KOI: 75.92% missing values
P. Min Mass (EU): 70.37% missing values
P. Mass (EU): 0.8500000000000001% missing values
P. Max Mass (EU): 100.0% missing values
P. Radius (EU): 0.31% missing values
P. Density (EU): 1.06% missing values
P. Gravity (EU): 1.06% missing values
P. Esc Vel (EU): 1.06% missing values
P. Teq Min (K): 1.1900000000000002% missing values
P. Teq Mean (K): 1.1900000000000002% missing values
P. Teq Max (K): 1.1900000000000002% missing values
P. Ts Min (K): 54.86% missing values
P. Ts Mean (K): 54.86% missing values
P. Ts Max (K): 54.86% missing values
P. Surf Press (EU): 1.06% missing values
P. Mag: 1.4500000000000002% missing values
P. Appar Size (deg): 0.31% missing values
P. Period (days): 3.8699999999999997% missing values
P. Sem Major Axis (AU): 0.8999999999999999% missing values
P. Mean Distance (AU): 0.8999999999999999% missing values
P. Inclination (deg): 81.32000000000001% missing values
S. Mass (SU): 1.21% missing values
S. Radius (SU): 2.8899999999999997% missing values
S. Teff (K): 2.55% missing values
```

```
S. Luminosity (SU): 0.72% missing values
S. [Fe/H]: 44.85% missing values
S. Age (Gyrs): 40.08% missing values
S. Appar Mag: 23.97% missing values
S. Distance (pc): 30.630000000000003% missing values
S. Mag from Planet: 1.1900000000000002% missing values
S. Size from Planet (deg): 3.3000000000000003% missing values
S. Hab Zone Min (AU): 2.97% missing values
S. Hab Zone Max (AU): 2.97% missing values
P. HZD: 1.1900000000000002% missing values
P. HZC: 1.06% missing values
P. HZA: 2.19% missing values
P. HZI: 2.19% missing values
P. SPH: 53.52% missing values
P. ESI: 1.24% missing values
```

c. Cardinality of categorical variables:

Cardinality of categorical variable was calculated so that accordingly desired algorithm can be selected to convert categorical variable to numerical by using conversion techniques like:

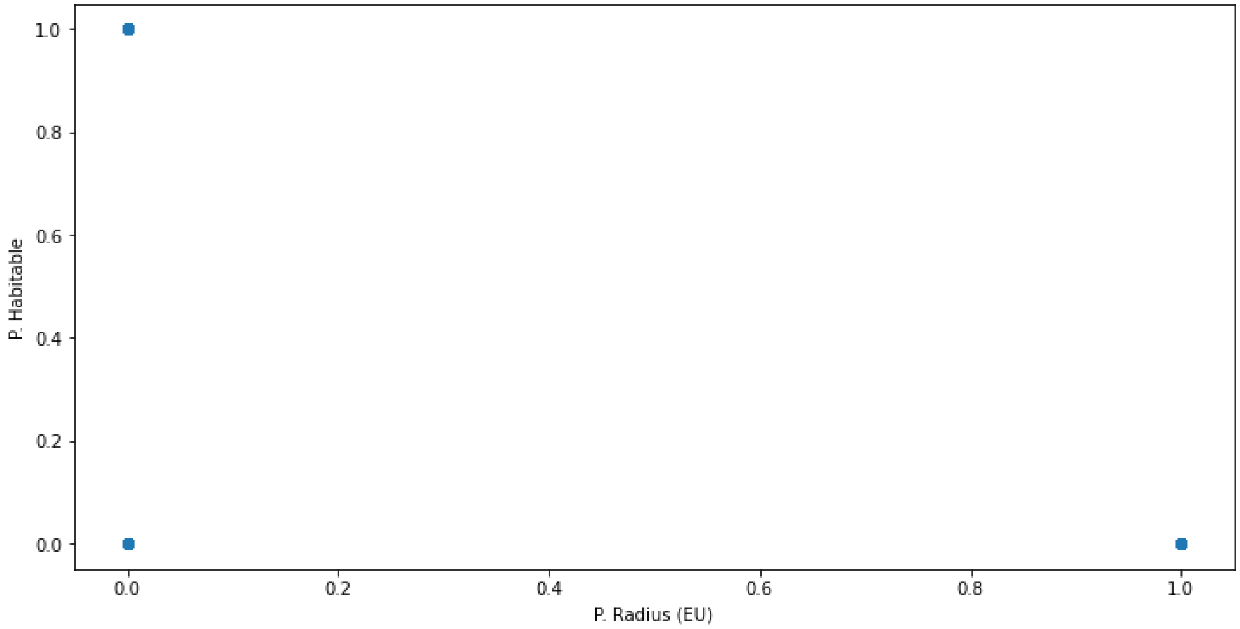
One-Hot Encoding

Label Encoder

d. Relationship between dependent and independent variables:

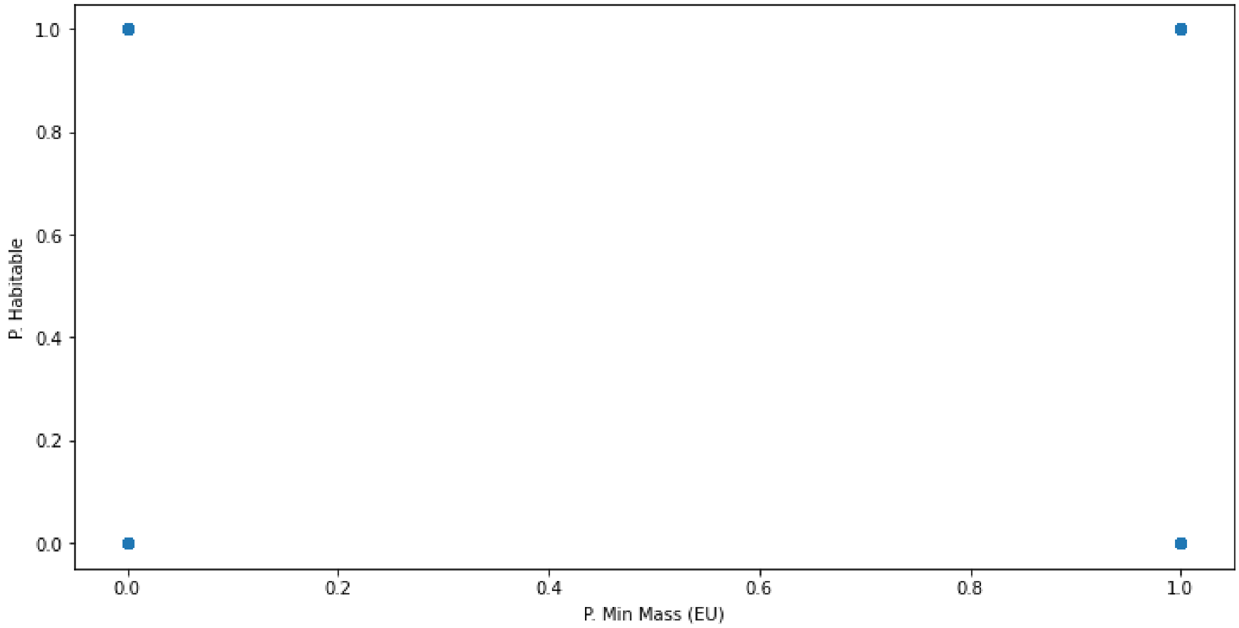
In this step a relationship between dependent and independent variable was observed based on the missing values.

For example:



Firstly, all the missing values were replaced by the value 1 and non-missing datapoints by 0.

Now in the above graph the relationship between Planetary Radius datapoints with missing values and non-missing values with the habitability of the planet is visualized. And the observation shows that there are no missing values in Planetary radius feature set that is contributing to the habitability of the planet positively i.e., none of the missing value datapoint is a habitable planet.



However, Planetary Minimum Mass feature set is accounting to the habitability of a planet positively by even the missing values i.e., missing value datapoints are also showing positive habitability index.

2. **Handling the missing values:**

   There are two primary ways of handling the missing values:

   1. **Deleting the missing values:**

      This can be achieved by either deleting the entire column or entire row consisting of missing values.

      If more than 50% of the datapoints are missing for a particular feature set than we can consider deleting entire column.

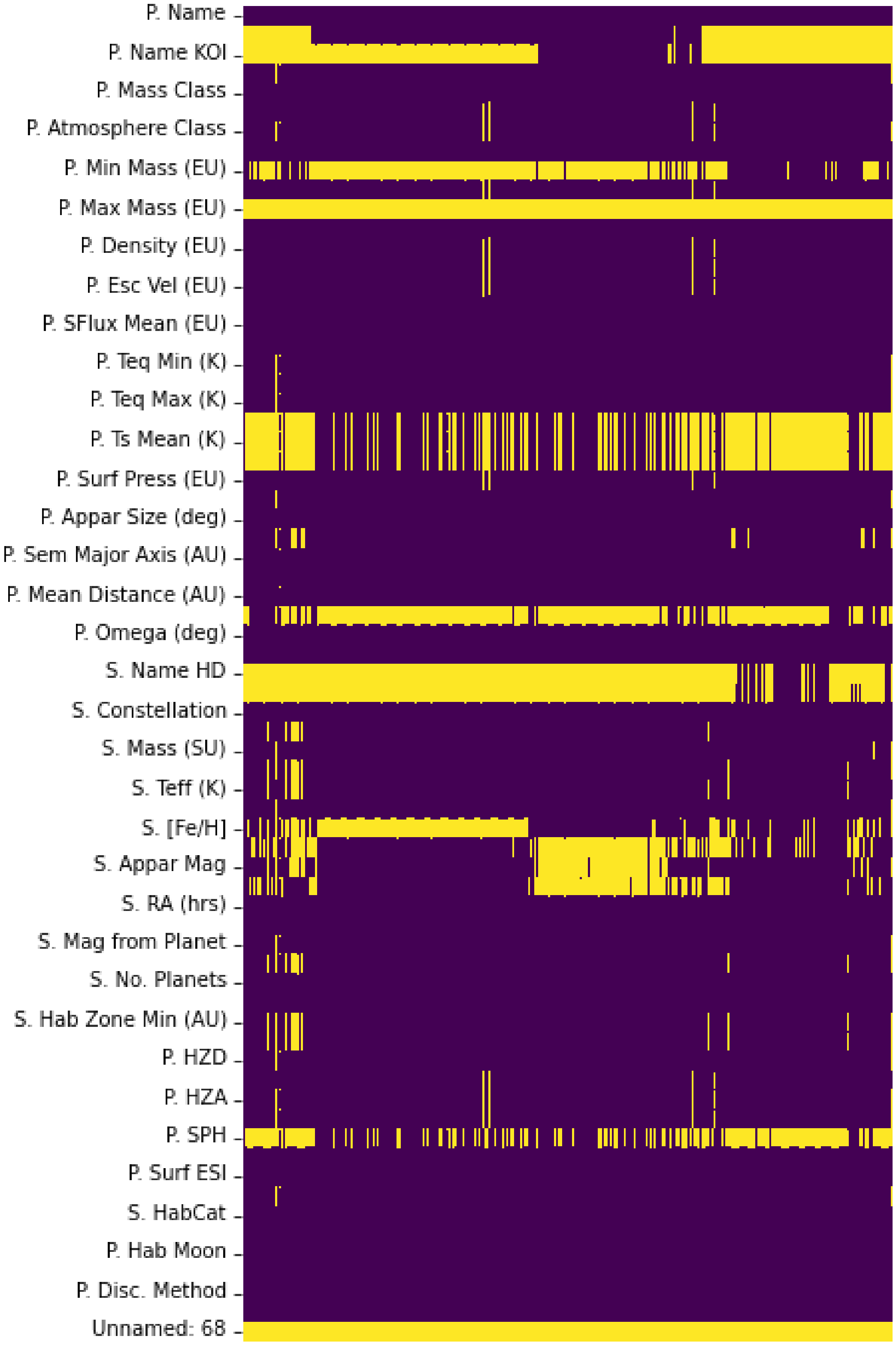
      However, if the number of missing values is significantly less than we can delete the entire row consisting of missing datapoints.

      These two are naïve methods of handling the missing datapoints.

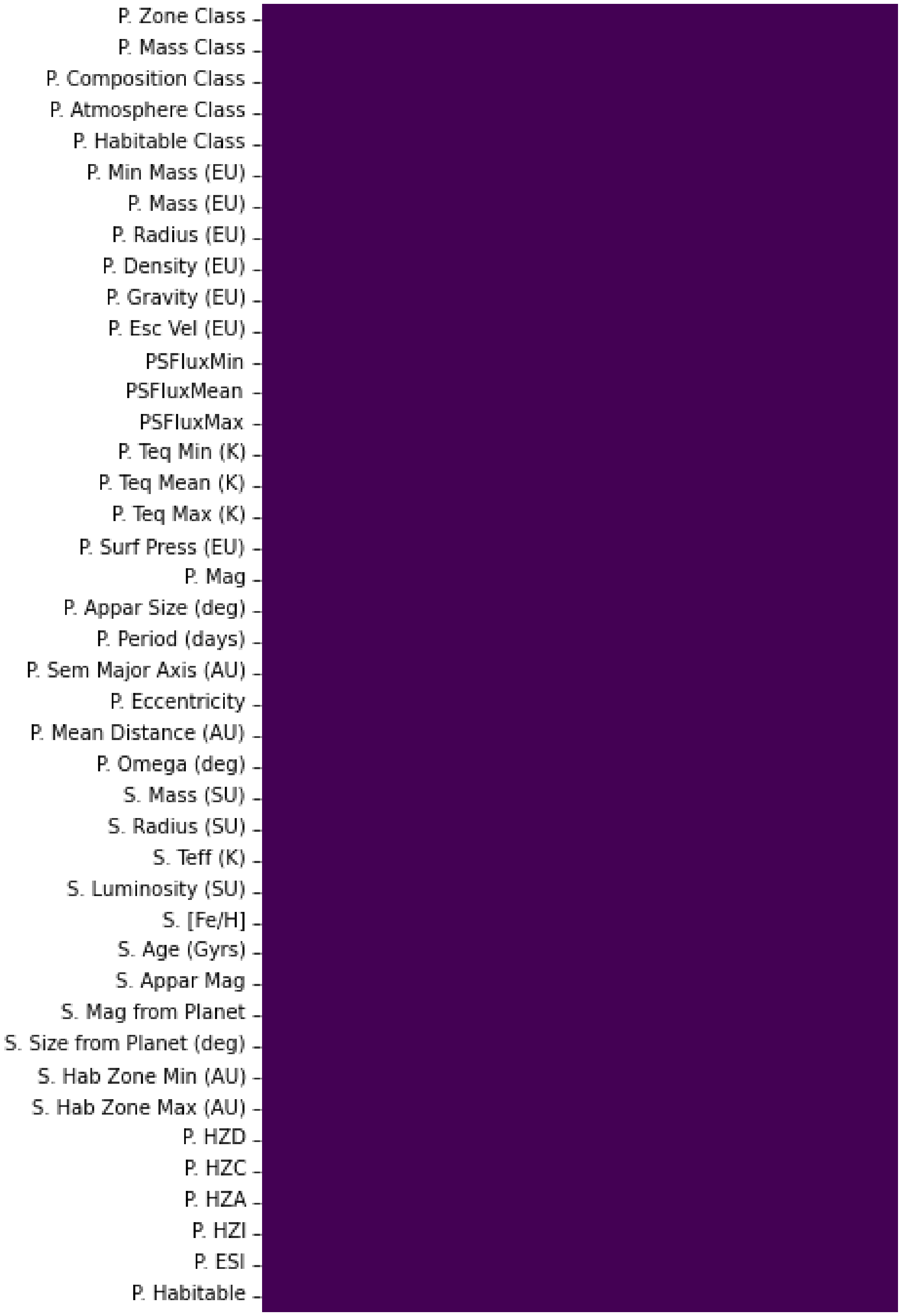
   2. **Imputing missing values:**

      This method includes replacing the missing values with some other value namely Mean or Average of the entire feature set. Of the above two techniques Mean is mainly preferred over Average because mean has a certain property that even if the datapoints with mean values are added mean of that set does not change.

Visualization of the dataset before dealing with missing values using heatmap function of the seaborn library is like:

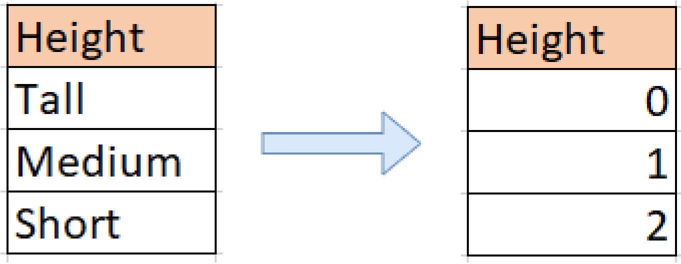


After dealing with missing values i.e., replacing missing values the visualization of the dataset changes to:



3. **Converting categorical variables to numerical variable:**
   We need to convert categorical variables to numerical variable so that our dataset with categorical feature set can be trained on a model that will only deal with numerical datapoints.
   For this purpose, label encoding was used instead of One-hot encoding algorithm because one hot encoding algorithm significantly increases the column numbers based on the cardinality of the feature set.
   For example, if the cardinality is 3 then number of columns that will be added is 3 which will increase the data and it will require more computational power to train our model.
   So, instead we used label encoding algorithm od deep learning which replaces the values using the range.
   For example, if cardinality of a feature set is 3 then they will be replaced within the column by the numbers in the range of 0 to 2 (0 to n-1).

   For example:



Height
Tall
Medium
Short

Height
0
1
2

   Now our dataset is ready to be fed into the deep learning model for training, validation and testing.

Result:

In the confusion matrix of different models:

0 stands for Psychroplanet

1 stands for non-habitable planet

2 stands for Mesoplanet

Mesoplanet as well as Psychroplanet both are considered as habitable planets.

Model 1:

3 hidden layers are used along with 1 input layer with 41 nodes and one output layer with 3 nodes.

Hidden layer 1: 20 nodes

Hidden layer 2: 12 nodes

Hidden layer 3: 12 nodes

3 dropout layers used to overcome the problem of model overfitting.

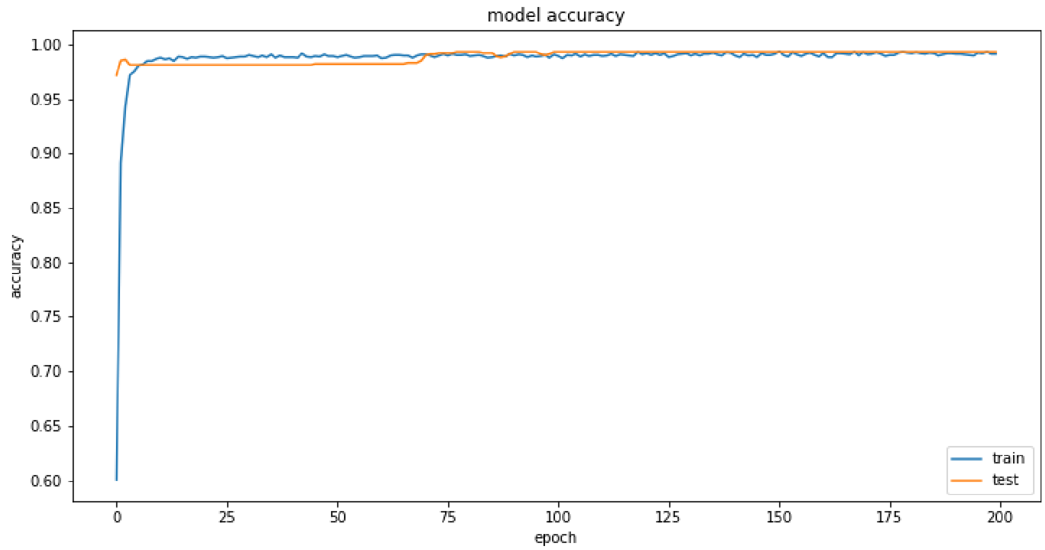
Activation function used is ReLU for hidden layers and softmax for output layer.

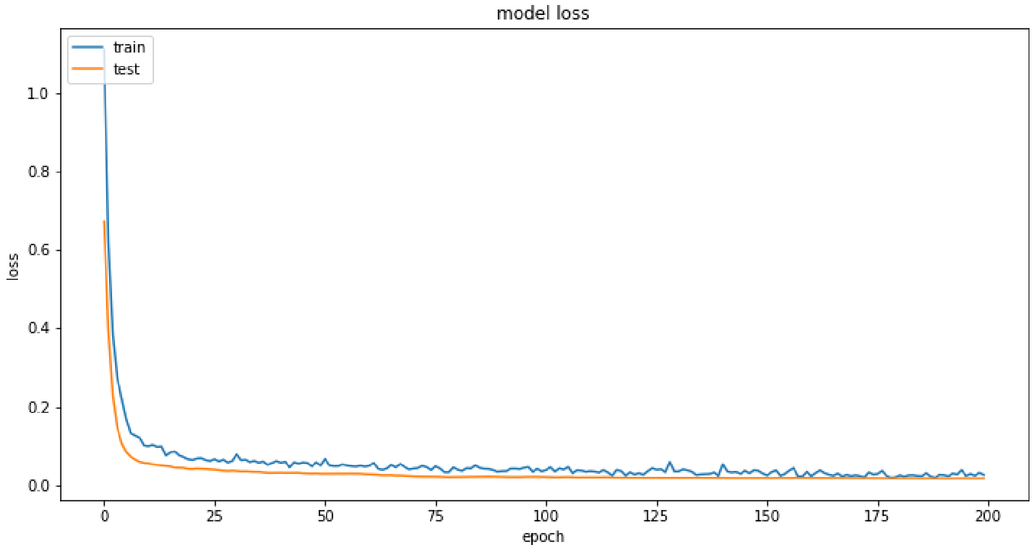
Optimiser used: Adamax

Batch size for training: 10

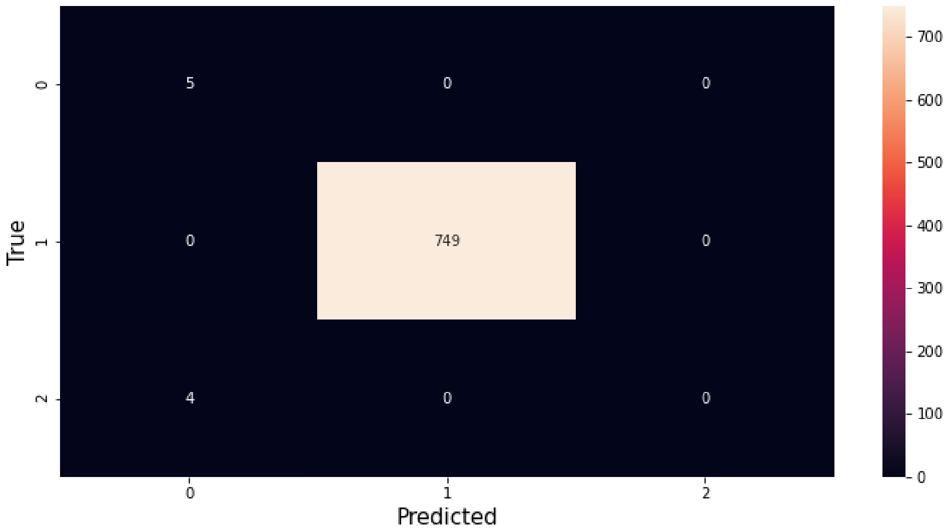
Epochs: 200

Accuracy score: 99.11





Heatmap of confusion matrix:



Model 2:

Activation function used is LeakyReLU for hidden layers and softmax for output layer.

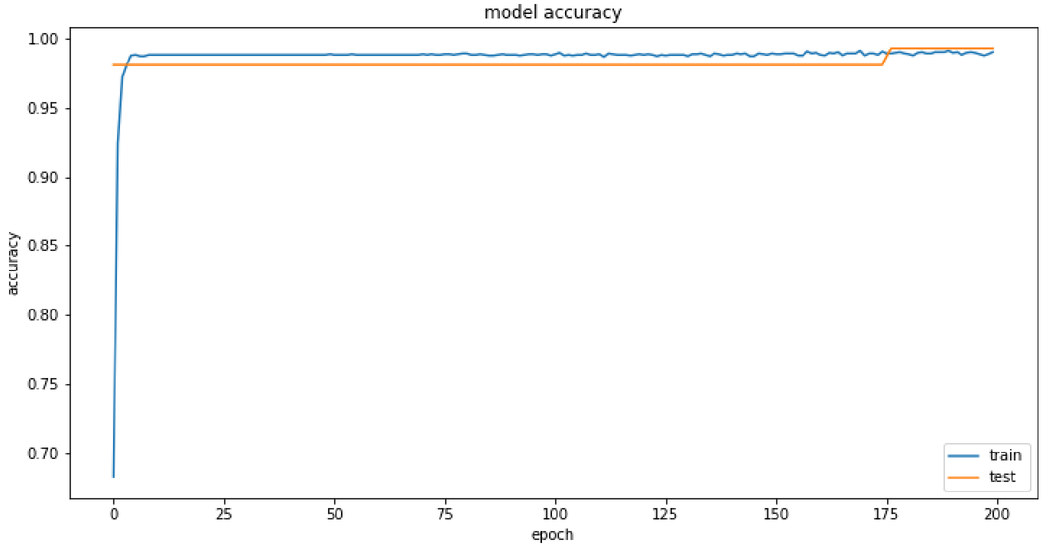
Alpha value: 0.5

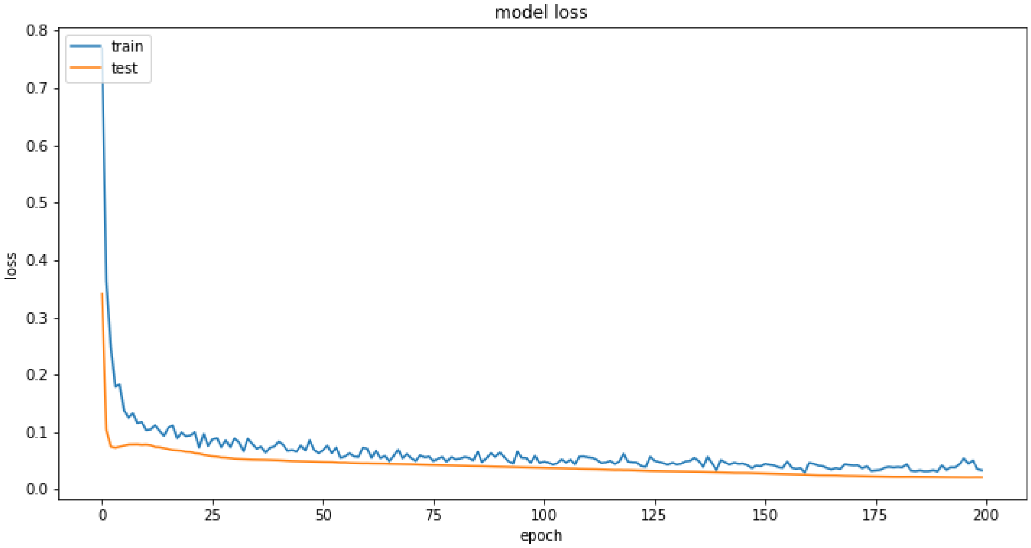
Optimiser used: Adamax

Batch size for training: 10

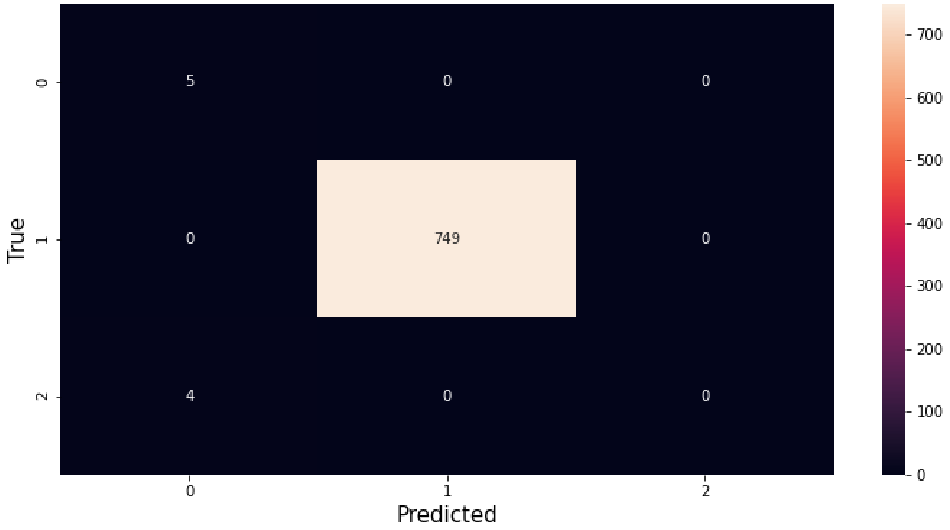
Epochs: 100

Accuracy score: 99.06





Heatmap of confusion matrix:



Model 3:

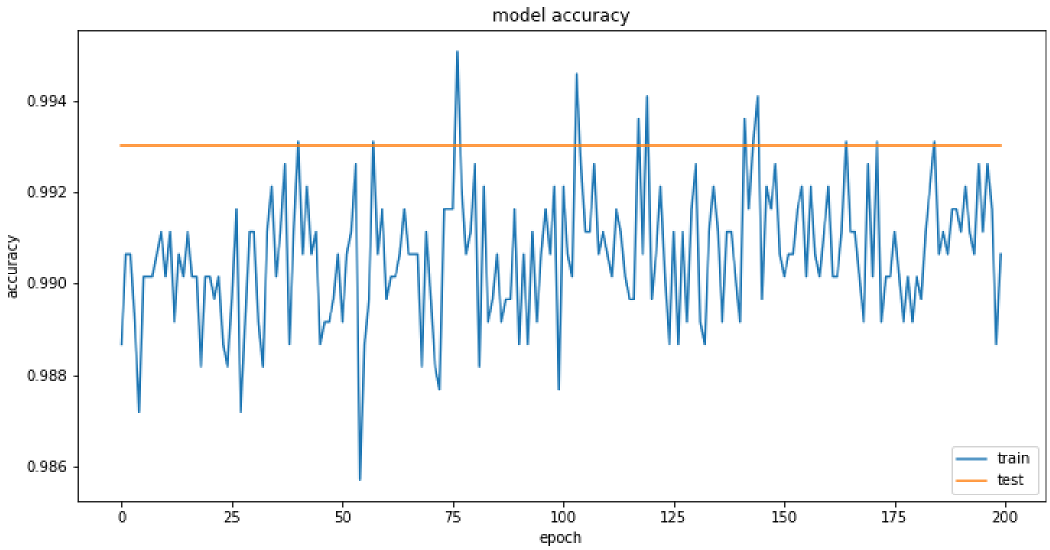
Activation function used is selu for hidden layers and softmax for output layer.

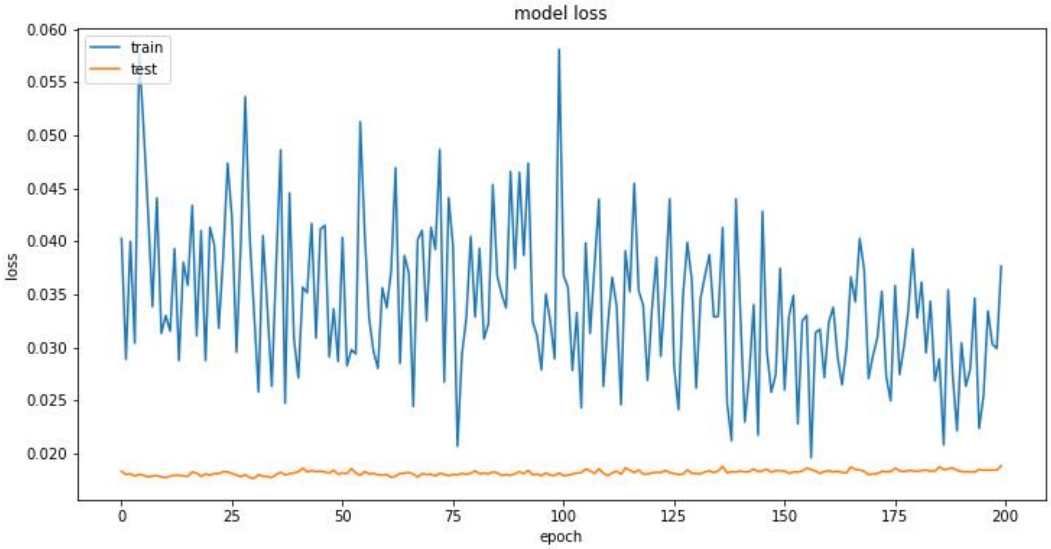
Optimiser used: Adamax

Batch size for training: 10

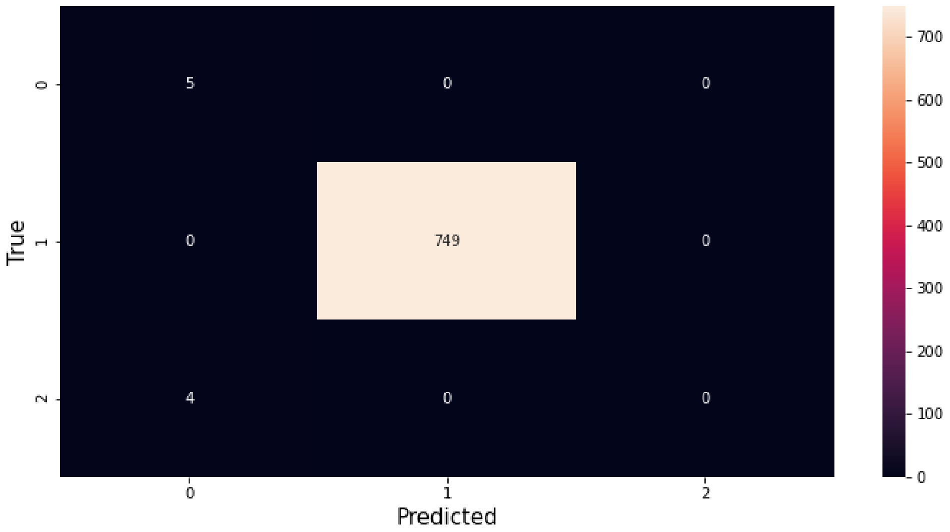
Epochs: 100

Accuracy score: 99.06





Heatmap of confusion matrix:



Accuracy score using LeakyReLU:

Accuracy score: 99.47

F1 Score using LeakyRelu:

F1 score: `[0.71428571, 1.        , 0.        ]`

Precision score using LeakyReLU:

Precision score: `[0.55555556, 1.        , 0.        ]`

Conclusion:

A working deep learning (Artificial Neural Network) model will be successfully created with advanced and never used before activation functions like SoftMax, Swish and Softplus and advanced optimizers like ADAGRAD and ADAM which will give much better accuracy in terms of prediction.

And also, the availability has sufficiently increased in the mean time which will also positively affect the accuracy and efficiency of our proposed model.

REFERENCES:

1. Mishra, Rajeev. "Predicting habitable exoplanets from NASA's Kepler mission data using Machine Learning." (2017).
2. Hora K. (2018) Classifying Exoplanets as Potentially Habitable Using Machine Learning. In: Saini A., Nayak A., Vyas R. (eds) ICT Based Innovations. Advances in Intelligent Systems and Computing, vol 653. Springer, Singapore.
3. Basak, S., Saha, S., Mathur, A., Bora, K., Makhija, S., Safonova, M. and Agrawal, S., 2020. CEESA meets machine learning: A Constant Elasticity Earth Similarity Approach to habitability and classification of exoplanets. *Astronomy and Computing*, *30*, p.100335.
4. Singh, S. P. . and Misra, D. K. . (2020) "Exoplanet Hunting in Deep Space with Machine Learning", *International Journal of Research in Engineering, Science and Management*, 3(9), pp. 187–192. Available at: http://journals.resaim.com/ijresm/article/view/323 (Accessed: 26 February 2022).
5. Jagtap, R., Inamdar, U., Dere, S., Fatima, M. and Shardoor, N.B., 2021, April. Habitability of Exoplanets using Deep Learning. In *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)* (pp. 1-6). IEEE.
6. Mathur, S., Sizon, S. and Goel, N., 2021. Identifying Exoplanets Using Deep Learning and Predicting Their Likelihood of Habitability. In *Advances in Machine Learning and Computational Intelligence* (pp. 369-379). Springer, Singapore.
7. Basak, S., Mathur, A., Theophilus, A.J. *et al.* Habitability classification of exoplanets: a machine learning insight. *Eur. Phys. J. Spec. Top.* **230,** 2221–2251 (2021).
8. Priyadarshini, I., Puri, V. A convolutional neural network (CNN) based ensemble model for exoplanet detection. *Earth Sci Inform* **14,** 735–747 (2021).
9. Rahman, M. and Afrin, N., 2018. *Finding habitable exo planets using boosting algorithm* (Doctoral dissertation, Brac University).
10. Saha, S., Nagaraj, N., Mathur, A. *et al.* Evolution of novel activation functions in neural network training for astronomy data: habitability classification of exoplanets. *Eur. Phys. J. Spec. Top.* **229,** 2629–2738 (2020).
11. Mousavi-Sadr, M., Gozaliasl, G., & Jassur, D. (2021). Exoplanets prediction in multiplanetary systems. *Publications of the Astronomical Society of Australia*, *38*, E015. doi:10.1017/pasa.2021.9