Interstellar Explorer

Rudri Jani Dhaval Deshkar Riya Shah Supan Shah

***Abstract*— The task is to classify the possible exoplanets from the database retrived from the NASA's Kepler Mission. Various Machine Learning approaches are used for classification of the given data. In initial phase, having been worked on pre-processing the data source and use the feature extraction algorithms such as Logistic Regression(LR) and Support Vector Machine(SVM).**

Keywords– Classification of exoplanets, Kepler Mission, Logistic Regression, Support Vector Machine, Machine Learning, Confusion Matrix

I. INTRODUCTION

An exoplanet is any planet far away from our solar system.The first exoplanet was discovered in the early 1990s. The first exoplanet was burst in the world stage was named 51 Pegasi b, Which was like "hot Jupiter" orbiting a Sun-like star far from 50 light-years. 1995 year was watershed year. Since then we have discovered thousands and more. The discovery of many of the exoplanets are so far off in a relatively small region of our galaxy, the Milky Way. We recognize from NASA's Kepler Space Telescope that there are planets more than stars in the galaxy. The size and mass plays a crucial role in determination of planets type. There are different varieties within the size/mass classifications. It is possible that critical size in planet formation: Planets that reach the size quickly attract thick atmospheres of hydrogen and helium gas, and balloon up into gaseous planets, while planets are smaller than this limit are not large enough to hold such an atmosphere and remain primarily rocky and terrestrial bodies.The planets which are in smaller size have that orbits close to their stars could be the cores of Neptune-like worlds that had their atmospheres stripped away.[1]

In Machine learning, classification describes the prdective modeling where a class label is predicted for a given input data. Based on this idea, there are two major techniques that is used for this task: Linear Regression(LR) and Support Vector machine(SVM). Logistic Regression algorithm is used for solving the classification problems with using the statistical approach. Support Vector Machine(SVM) is used for both classification and regression problems which is based on geometrical properties of the data. This powerful techniques are efficient to determine weather it is the exoplanet or not. Fig.1 shows the random picture of exoplpanets.

Fig. 1. Exoplanets

II. LITERATURE SURVEY

The lack of to confirm the individual planet candidates by statistical techniques should not be examined as a failure of the Kepler mission (or of the statistical techniques themselves, which are applicable when their underlying hypothesis remain valid). Kepler's aim was to establish the frequency of Earth-like planets in the Galaxy, and the judgement that at least 2% (and possibly as many as 25%) of stars host an Earth-size planet in the shorter-period range of 50–300 days (Burke et al. 2015)is not strongly affected by this result.The reliability of the Kepler sample in this regime is much higher, and sharpness of occurrence rate calculations is less sensitive to catalog accuracy.

It will be required to account for record accuracy to properly extend occurrence rate calculations out toward the longer periods that encompass the habitable zone around G-type stars.[6]

III. IMPLEMENTATION

The project will use classical machine learning methods for the classification of planets as exoplanets or not using features extracted by the Kepler space telescope. Importing the necessary libraries for the Machine learning algorithm and working on Pre-proecessing the data.

A. Pre-processing

In the phase of data pre-processing, The data is retrived from the NASA's Exoplanet Database.[3] Data having 8.481% of the data was null. In the given dataset, An exoplanet candidate is a probable planet found by a telescope however has not yet been demonstrated to really exist or not. It is possible for some candidates to turn out to be "false positives."[2] A planet is only considered "confirmed" once it is verified through additional observation using two other telescopes. There are so many planets candidates awaiting confirmation. But time on telescopes is considered a precious resource and it takes a lot of computing time to search the determination of the targets. This one area is where amateur scientists can work with NASA data to help filter the targets and even discover exoplanets. Where computers might miss a single transit, humans can detect little brightness dips in data that might tell us there is a planet to be found. After finding irrelevant columns from dataset it is removed.[5] The missing values of the all the rows is removed. After that data is normalized. For the dataset, hot encoding is necessary for columns having categorical values. Hence, One hot encoding of categorical data is done.[4] After that it is creating a new column on the basis of candidate and confirmed planets which is the target values. The data is splitted into train and test data in 80-20% and applied on the model. The following chart shows the flow of the task,

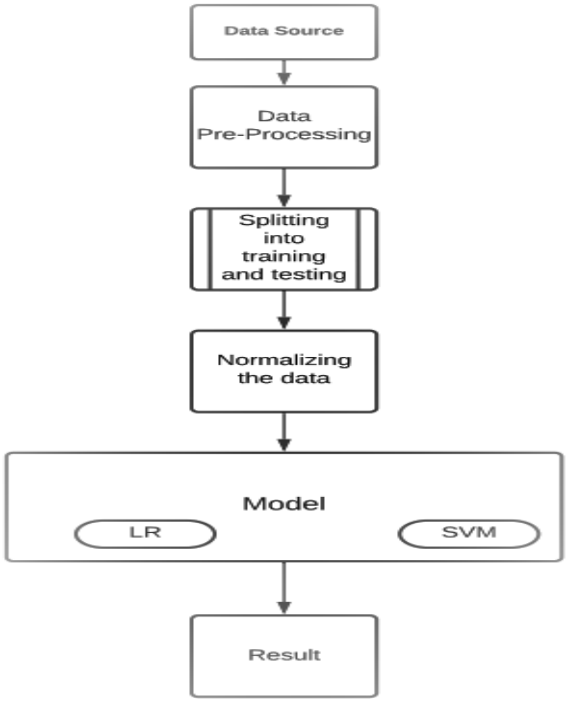

Fig. 2. Flow chart

B. Logistic Regression

Logistic Regression is significant approach of regression analysis to conduct when the dependent variables is binary. It is Predictive analysis. It is used to describe the data and to explain the relationship between one dependent binary variable and independent variables. The equation of Logistic regression :

$$Y = \exp(b0 + b1 * X)/(1 + \exp(b0 + b1 * X))$$

In this equation Y is expected output, b0 is bias and b1 is it's multiplication factor for the input X.

We are using Sigmoid Function, Cost Function and Gradient Descent for the Logistic regression

Sigmoid Function:

$$\sigma(x) = 1/(1 + \exp(-x)$$

Cost Function:

$$J = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log(a^{(i)}) + (1 - y^{(i)}) \log(1 - a^{(i)})$$

Gradient Descent:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} f(\theta_0, \theta_1) \text{ for j=0,1}$$

Output: It will gives the smooth curve.

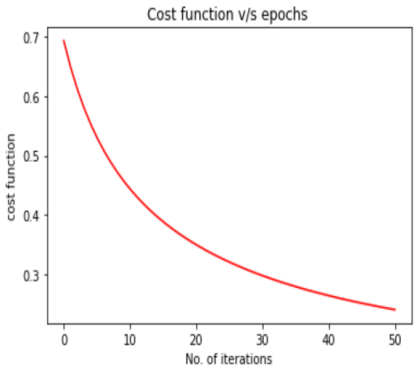

Fig. 3. Output: Cost Function

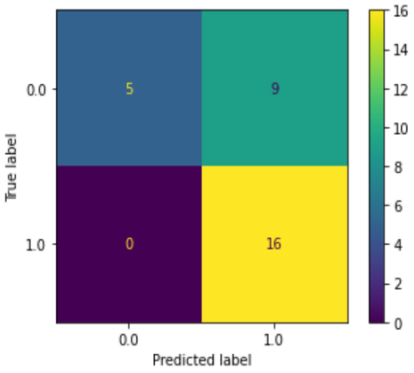

Fig. 5. Confusion Matrix Of SVM

C. Support Vector Machine

Support vector machine (SVM) is powerful and flexible class of supervised algorithms for both classification and regression.

IV. RESULTS

After fitting the training data in different models and predict the test datasets. The first model used is logistic regression to classify the exoplanets using normal equation and we find the F1 Score is 93.75%. Another model used is SVM and we find the F1 Score is 78%. Here,Confusion Matrix shows the results of the task

Results:

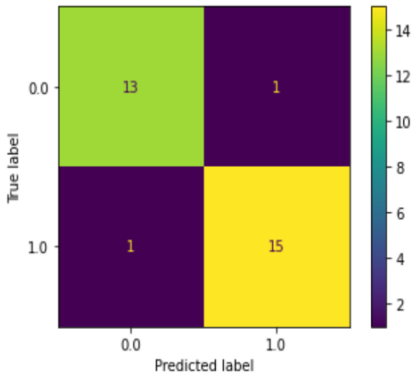

Fig. 4. Confusion Matrix Of LR

V. CONCLUSION

Hence, the result shows the accuracy of the two different Machine learning techniques named as Logistic Regression(LR) and Support Vector Machine(SVM). Which generates the classification report and confusion report for unseen data and classify the exoplanets whether it is confirmed or not.

As there are many correlations between columns, using PCA, the project will try to minimize the dimensions and compress the data.

We will create dashboard where user can key-in feature values and get instant predictions.

REFERENCES

[1] https://exoplanets.nasa.gov/
[2] https://exoplanetarchive.ipac.caltech.edu/docs/program.interfaces.html
[3] https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTblsconfig=cumulative
[4] https://exoplanetarchive.ipac.caltech.edu/docs/API.kepcandidate.columns.htm
[5] https://exoplanets.nasa.gov/what-is-an-exoplanet/overview/
[6] https://iopscience.iop.org/article/10.3847/1538-3881/aabae3/meta