**PES University, Bengaluru**

**V Sem**

**Session: August – December, 2019**

**UE17CS303 – Machine Learning Assignment**

PROJECT REPORT

On

“EXOPLANET CLASSIFICATION”

Submitted by

**ATHIRA A D** PES1201701640

**NIVEDITHA C U** PES1201701633

**S MANEESHA** PES1201700024

**Department of Computer Science & Engineering**

**PES University**

(Established under Karnataka Act No. 16 of 2013)

100 Feet Ring Road, Bengaluru – 560 085, Karnataka, India

**PROBLEM STATEMENT:** Exoplanet Classification

**About the dataset:**

* Exoplanets – planets that are outside our solar system, also known as extra-solar planets or exo-planets.
* The dataset consists of a total of 68 features, 13 categorical and the remaining 55 are continuous.
* It constitutes 3876 rows where each row represents an exoplanet belonging to anyone of the below listed habitable class:

|  |  |  |
| --- | --- | --- |
| Habitable Class | # of planets in each class | Description |
| Non-Habitable | 3820 | 1. Planets, mostly too hot or too cold, may be gaseous, with non-rocky surfaces. 2. Such conditions don’t favor habitability. |
| Mesoplanet | 31 | 1. Earth-like planets, they have sizes between that of Ceres (the largest minor planet in our solar system) and Mercury. 2. The average global surface temperature - between 0◦C and 50◦C. 3. Earth-similarity is no guarantee of habitability. |
| Psychroplanet | 18 | 1. Mean global surface temperature - between -50◦C to 0◦C. 2. Temperatures are colder than optimal for sustenance of terrestrial life. 3. Some psychroplanets are still considered as potentially habitable candidates. |
| Hypopsychroplanet | 3 | - |
| Thermoplanet | 3 | 1. A habitable planet with a surface temperature of 50–100°C. 2. Warmer than is optimal for most terrestrial life. |

**System Architecture:**

**PREPROCESSING**

**SAMPLING**

**TRAIN DATA**

Testing the Model

**TEST DATA**

Training the Model

**TRAINED MODEL**

**OUTPUT**

**ML Techniques employed:**

* **Preprocessing:**
* **Dimensionality Reduction:**

1. Since majority of the columns have a greater number of missing values, we removed such columns which have missing value ratio > 90%.
2. Based on the variance also few columns were removed. This is because an attribute with low variance do not really separate datapoints in any way so it doesn’t help for classification.
3. Some of the columns like Name of the Host star, Year of Discovery are unnecessary, such columns were also removed.

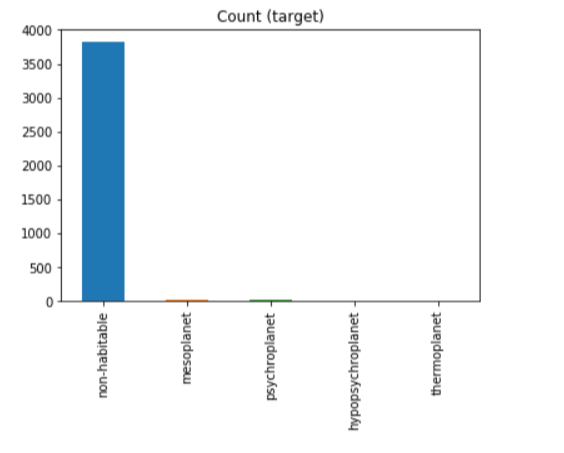
* **Categorical Encoding:**

1. Our dataset contains categorical columns like P. Mass Class, P. Atmospheric Class, P. Habitable Class, we applied self-developed label encoding algorithm to convert categorical columns into numerical ones since ML model does not accept string values.

* **Replacing Missing Values:**

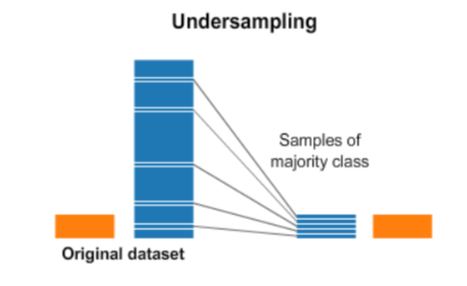
1. We replace non-categorical columns with class-wise mean and categorical columns with mode.

* **Sampling:**



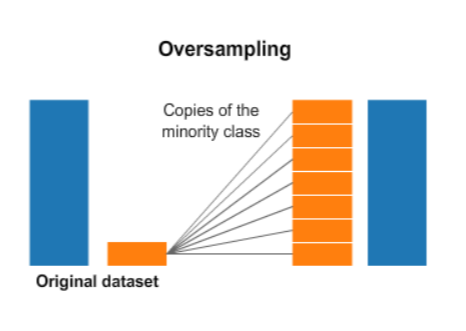
From the above Figure, we can understand that our dataset is highly imbalanced. Hence, we apply 2 types of sampling techniques:

1. Undersampling:



For undersampling, we removed rows containing ‘gas’ as composition class since all the planets with gas composition is non-habitable.

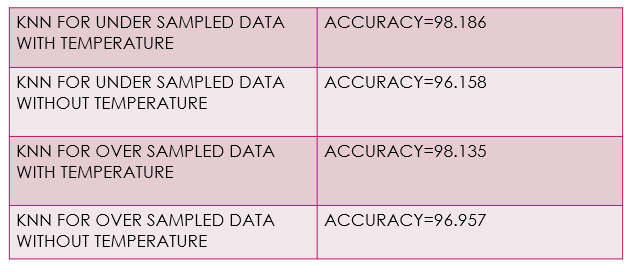
1. Oversampling:



For oversampling, we duplicated random records from minority class.

* **Classification:**
* We applied KNN (K-Nearest Neighbor) with k=3.
* It has already been known that surface temperature can distinguish habitable planets from non-habitable ones with large degree of precision so it is necessary to verify whether our classifier classifies properly even without this feature. Hence, we tried KNN with and without Temperature attribute.

**Summary of result:**



**Conclusions:**

* The classifier gave high accuracy even without Temperature attribute.
* Even though sampling helped us to balance the class, from the analysis we understood that these techniques also have their own weaknesses.
* Since oversampling just duplicates random records from minority class, this can cause overfitting.
* In undersampling, since we removed planets with gaseous composition this may cause loss of information.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*