DAL Assignment 6: A Mathematical Essay On Support Vector Machine

C Vamshi Krishna Reddy
CH20B112
EE5708: Data Analytics Laboratory
IIT Madras, TamilNadu, India
ch20b112@smail.iitm.ac.in

***Abstract*—This document serves as a model for showcasing the application of the Support Vector Machine to build a model that helps us to classify a star is a pulsar star or not based on eight different parameters. The first four are simple statistics obtained from the integrated pulse profile (folded profile). The remaining four variables are similarly obtained from the DM-SNR curve.**

***Index Terms*—Classification, Support Vector Machine, Statistical Modeling, Supervised Learning**

I. Introduction

We all know that classification problems are one of the primary tasks solved with the help of Machine learning apart from regression. These algorithms create patterns between input and output variables using already existing data and classify the target variable based on the new input data provided. There are 3 types of machine learning algorithms: supervised, unsupervised, and reinforcement learning. In this paper, we are using Support Vector Machine(SVM), which comes under the supervised category.

In simple terms, the classifier is defined as a machine learning model which is used to discriminate different labels based on certain input features. In an SVM classifier, a hyperplane is constructed between the two classes. In other words, the data points on one side of the hyperplane will all be assigned to one category, while the data points on the other side of the line will be assigned to a different category. This implies that the number of possible hyperplanes is unlimited. SVM selects the hyperplane that divides the data and is as far from the nearest data points as it may be making it better than some of the other algorithms, like k-nearest neighbors.

This paper aims to use the Support Vector Machine classifier in explaining how parameters like mean, standard deviation, excess kurtosis, and Skewness of integrated profile and DM-SNR curve of a star determine if it is a pulsar or not.

The paper provides a detailed overview of the SVM classifier by solving a real-world problem and gives insights into the relationship between the Target variable and the above-mentioned independent input variables.

II. Support Vector Machine

Classification is a widely employed application in machine learning, especially in supervised settings. It essentially helps us understand how certain independent factors determine which class it belongs to. In simpler terms, it's like a tool used in machine learning to make classifications. The core idea is to teach algorithms how different independent variables are linked to an outcome or dependent variable. Once the model grasps this relationship, it becomes capable of predicting what might happen when faced with new, unexpected data or completing missing parts in existing data.

Support Vector Machine, or SVM, can be used to solve both Classification and Regression problems. However, it is largely employed in Machine Learning Classification issues. The SVM algorithm's objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future. A hyperplane is a name given to this optimal decision boundary. Different hyperplanes might be used to split the two classes of data points. Finding a plane with the greatest margin that is, the greatest separation between data points from both classes is our goal. Maximizing the margin distance adds some support, increasing the confidence with which future data points can be categorized. SVM selects the extreme vectors and points that aid in the creation of the hyperplane. Support vectors, which are used to represent these extreme instances, form the basis for the SVM method. Take a look at the diagram below, where a decision boundary or hyperplane is used to categorize two distinct categories.

Based on the type of hyperplane of the algorithm, the Support Vector Machine classifier is classified into two subgroups: Linear SVM, and Non-linear (or) Kernel SVM.

•**Linear SVM:** When data are linearly separable or when categorizing a dataset into two categories using a single straight line, linear SVM is utilized. These data points are known as linearly separable data, and the utilized classifier is referred to as a linear SVM classifier.

•**Non-Linear SVM:** Non-Linear SVM is used for data that cannot be classified using a straight line because they are non-linearly separable. To achieve this, we employ a technique called the kernel trick, which places data points in a higher dimension where they can be divided using planes or other mathematical operations. These data points are known as non-linear data, and the utilized classifier is known as a non-linear SVM classifier.

A. Cost Function

The optimal SVM classifier is obtained by solving a langrage problem shown below

$$\min_{w,b}\left(0.5||w||^2+\sum_{i=1}^{m}\alpha_i|(y_i(w.x+b)-1)|\right) \qquad (1)$$

subject to $\alpha_i \geq 0,$ for $i = 1, 2, ..., m$

Here alpha refers to the Langrage multipliers or KKT multipliers to be more accurate as we are dealing with inequality constraints. In this model, we have decided to use a polynomial kernel for classification as the provided data is non-linearly separable.

B. Evaluation Model

In classification models like Support Vector Machine, confusion matrix, and f1 score are used to evaluate the performance of the model. Here since our problem has 2 output classes we get 2 rows for each parameter, As a side note, F1-Score is defined as the harmonic mean of recall and precision values of the classification problem. The formula for F1-Score is as follows:

$$F1 = \frac{2 * precision * recall}{precision + recall} \qquad (1)$$

III. The Problem

The objective is to develop a model to decide whether a star is pulsar or not based on factors like Mean of the integrated profile, standard deviation of the integrated profile, Excess kurtosis of the integrated profile, Skewness of the integrated profile, Mean of the DM-SNR curve, Standard deviation of the DM-SNR curve, Excess kurtosis of the DM-SNR curve, Skewness of the DM-SNR curve as the input features. The process initiates with data cleaning and then exploratory data analysis where we predict whether the star is a pulsar or not.

A. Data Cleaning

The given dataset contains around 12528 rows and 9 columns. Upon looking at the information of data we see that there are missing values or nulls in the data set especially in “Excess kurtosis of the integrated profile”, “Standard deviation of the DM-SNR curve”, and “Excess kurtosis of the DM-SNR curve”. They should be either eliminated or modified using common methods such as mean estimation-based filling, distribution-based filling, etc. In the dataset, we observe that roughly 15 percent of the excess kurtosis of integrated profile data is missing. This proportion is likely small enough to drop those rows instead of adding a bias to the dataset by imputing them with various central tendencies like mean etc. Similarly other 2 columns also have less than 10 percent nulls so we opted to drop those rows.

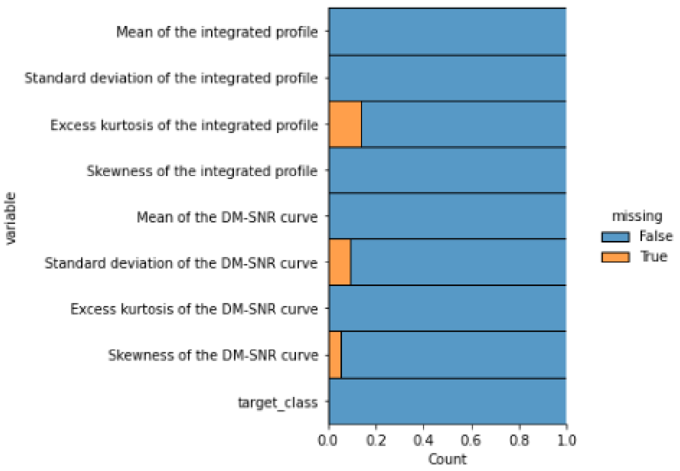


Fig. 1. Percentage Count of Null values.

B. Exploratory Data Analysis and Visualisations

The provided data set includes star information like mean of the integrated profile, standard deviation of the integrated profile, excess kurtosis of the integrated profile, skewness of the integrated profile, mean of the DM-SNR curve, standard deviation of the DM-SNR curve, excess kurtosis of the DM-SNR curve, skewness of the DM-SNR curve. We can see that there are 9 variables in the dataset. 8 are continuous variables and 1 is a discrete variable. The discrete variable is the target class variable. It is also the target variable. These details of the star are important and will be used to train a model that can predict whether the given star is pulsar or not, based on the “ground truth” i.e. known data. So our main motive for this project is to develop an SVM classification model using a training data set and then predict the Target label on unseen data.

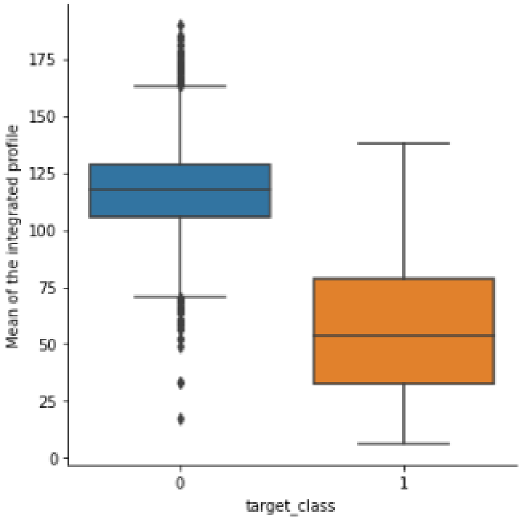


Fig. 2. Box plot of Mean of the Integrated Profile.

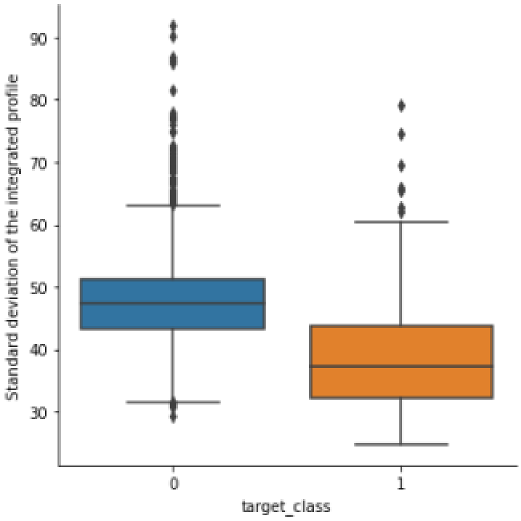


Fig. 3. Box plot of Standard Deviation of the Integrated Profile.

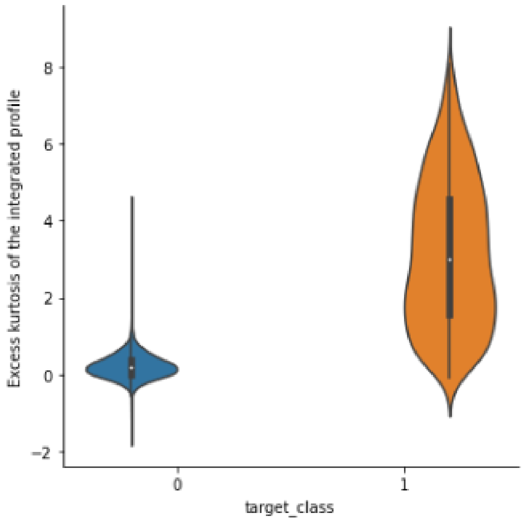


Fig. 4. Violin plot of the Excess kurtosis of the Integrated Profile.

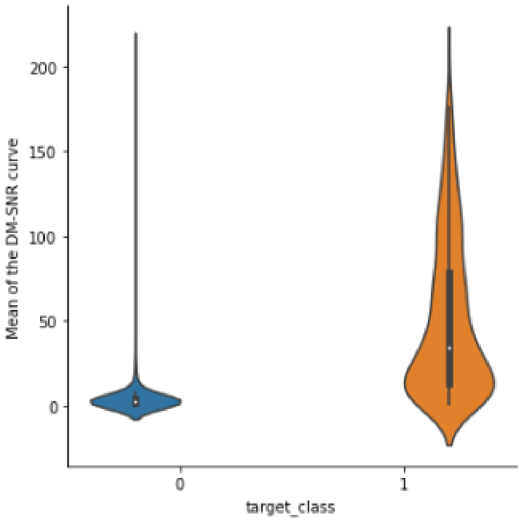


Fig. 5. Violin plot of Mean of the DM-SNR Curve.

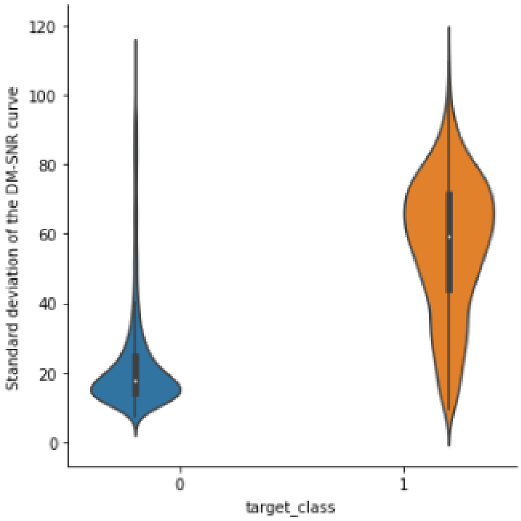


Fig. 6. Violin plot of the standard deviation of the DM-SNR Curve.

From the correlation heatmap, we can see that the mean and standard deviation of the integrated profile and excess kurtosis and skewness of the DM-SNR curve are negatively correlated with the target class. So larger values of these variables may potentially mean that the star doesn't belong to a pulsar, Similarly, other columns like excess kurtosis and skewness of the integrated profile and mean and standard deviation of the DM-SNR curve positively correlate with the data. So having larger values in these columns increases the probability of the star being pulsar. The violin plots and box plots plotted also the distribution of these variables for the Target variables being 0 and 1 (Pulsar star or not).

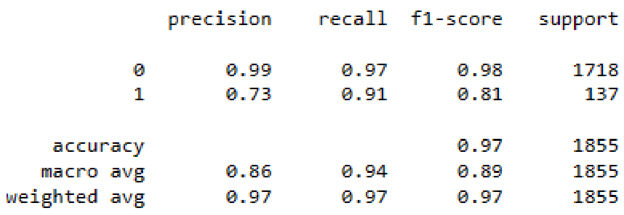
C. Model

The dataset contains outliers which will affect the model accuracy. So we need to remove the outliers before training the dataset. We can see that there are leading spaces (spaces at the start of the string name) in the column names. So, we will remove these leading spaces. Since the column names are very long. So, I will make them short by renaming them. IP stands for integrated profile and DM-SNR stands for delta modulation and signal-to-noise ratio. This pre-processed data is now divided into y referring to the target labels and X referring to input data. This data is split into 70 percent for training the model using a test train split from the sklearn model selection with a random state of 42. Now the SVM model is developed on this dataset with the 8 features as input and target class as output. We ran the SVM model using different kernels which gave the same accuracy.

Now we have to use our model to predict the test data set and find the F1 score using this predicted and true data.

We cannot say that our model is very good based on the accuracy **0.97**. We must compare it with the null accuracy. Null accuracy is the accuracy that could be achieved by always predicting the most frequent class. The null accuracy for this dataset is 0.9. We can see that our model accuracy score is 0.97 but the null accuracy score is 0.9. So, we can conclude that our SVM classifier is doing a very good job in predicting the class labels.

The classification report of the Support Vector Machine Classifier is given below. We achieved an accuracy of **0.97**.



```
              precision    recall  f1-score   support

           0       0.99      0.97      0.98      1718
           1       0.73      0.91      0.81       137

    accuracy                           0.97      1855
   macro avg       0.86      0.94      0.89      1855
weighted avg       0.97      0.97      0.97      1855
```

Fig. 7. Classification Report.

D. Insights and Observations

From the heat map of correlation, we can see that there is a positive correlation between excess kurtosis of the integrated profile, skewness of the integrated profile, mean of the DM-SNR curve, standard deviation of the DM-SNR curve, and Target class. So, larger values of these features will predict the star as pulsar star whereas smaller values will predict the star as not a pulsar. We can also observe a negative correlation between the mean of the integrated profile, standard deviation of the integrated profile, excess kurtosis of the DM-SNR curve, skewness of the DM-SNR curve, and target class. So, larger values of these features will predict the star as not a pulsar star whereas smaller values will predict the star as a pulsar star.

The "Confusion matrix" gives us a summary of correct and incorrect predictions broken down by each category. A "classification report" is another way to evaluate the performance of our classification model. It displays the precision, recall, f1, and support scores for the model.

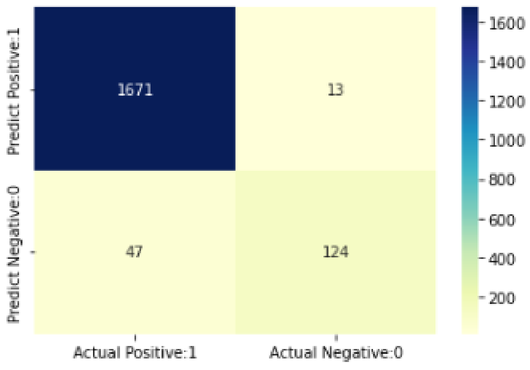
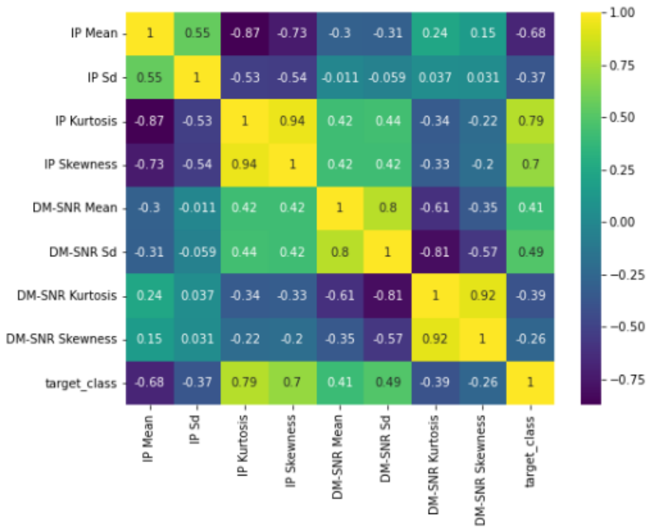


Fig. 8. Confusion Matrix.



	IP Mean	IP Sd	IP Kurtosis	IP Skewness	DM-SNR Mean	DM-SNR Sd	DM-SNR Kurtosis	DM-SNR Skewness	target_class
IP Mean	1	0.55	-0.87	-0.73	-0.3	-0.31	0.24	0.15	-0.68
IP Sd	0.55	1	-0.53	-0.54	-0.011	-0.059	0.037	0.031	-0.37
IP Kurtosis	-0.87	-0.53	1	0.94	0.42	0.44	-0.34	-0.22	0.79
IP Skewness	-0.73	-0.54	0.94	1	0.42	0.42	-0.33	-0.2	0.7
DM-SNR Mean	-0.3	-0.011	0.42	0.42	1	0.8	-0.61	-0.35	0.41
DM-SNR Sd	-0.31	-0.059	0.44	0.42	0.8	1	-0.81	-0.57	0.49
DM-SNR Kurtosis	0.24	0.037	-0.34	-0.33	-0.61	-0.81	1	0.92	-0.39
DM-SNR Skewness	0.15	0.031	-0.22	-0.2	-0.35	-0.57	0.92	1	-0.26
target_class	-0.68	-0.37	0.79	0.7	0.41	0.49	-0.39	-0.26	1

Fig. 9. Heat map of correlation.

IV. Conclusion

Rare neutron stars called pulsars emit radio waves that can be picked up on Earth. As sensors of space-time, the interstellar medium, and states of matter, they are of great scientific interest. To speed up analysis, pulsar candidates are now being automatically labeled using machine learning methods. Predicting whether a star is a pulsar start or not is the main task of these machine learning methods. Analysing the confusion matrix and classification report of our project we can say that our SVM Classifier model yields excellent performance in solving this problem. The accuracy of our model is 0.97 on the validation data. As our model performed extremely well on the validation data, we can now use this model on the testing data set. Here we didn't remove or modify any column name. So we can directly use this model on the test data set ny dropping the null values in the target class of test data set.

References

[1] Bishop, Christopher M., Pattern Recognition and Machine Learning, 2006.

[2] https://seaborn.pydata.org/tutorial/categorical.html.

[3] Aurelien Geron., Hands on Machine Learning with Scikit-Learn and Tensorflow, 2019.

[4] https://www.geeksforgeeks.org/support-vector-machine-algorithm/.

[5] https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/.

[6] https://towardsdatascience.com/the-f1-score-bec2bbc38aa6.