DAL 2023 Assignment 6: A Mathematical Essay on Support Vector Machines

Arun Palaniappan
Department of Mechanical Engineering
Indian Institute of Technology Madras
Chennai, India
me20b036@smail.iit.ac.in

***Abstract*—This paper provides an overview of the mathematical formulations and applications of Support Vector Machines. First, the fundamentals of Support Vector Machines are explored, as well as the specifics necessary to comprehend the research. The problem statement and data-set are introduced in a detailed way. The data-set under consideration is one of the prediction of pulse stars, to explore which attribute of the pulsar is more contributing to the model's prediction. To accomplish this efficiently, the data has been thoroughly evaluated, and both visual and quantitative insights have been provided in this paper. Finally, the results of the Support Vector Machine model are reported in a detailed manner. This paper has been improved in the "Problem" section by introducing Leave One Out Cross-Validation.**

***Index Terms*—Support Vector Machines, pulsar, F_1-score, precision, recall, mean, kurtosis, visualization**

I. Introduction

Support Vector Machine(SVM) is a powerful and versatile supervised learning algorithm primarily used for regression tasks. It is a parametric technique that aims to find the optimal hyperplane, which maximizes the margin between different classes in the feature space. The data points that lie closest to the decision surface are called support vectors, which define the hyperplane. SVM uses an intelligent kernel transformation of the feature space to maximize the margin. By using a variety of kernel functions, SVM is extremely effective for non-linear classification.

Essentially, the original data points are projected into a higher-dimensional space by the kernel functions, enabling the separation of data points that are not linearly separable in the original space. This facilitates the discovery of an optimal hyperplane in the transformed feature space where the margin, or the distance between support vectors to the hyperplane, can be maximized. To test how our model will perform on unforeseen data-sets in the future, we build a train-validation split, then train the model (minimize the impurity) on the train split, and then check the prediction accuracy on the validation split.

Pulsars, a unique class of neutron stars, are cosmic light-houses that emit beams of radiation with regularity and their pulses range from milliseconds to seconds. These stars possess intense magnetic fields that channel streams of particles from their magnetic poles, creating powerful beams of light. The DM-SNR curve represents the relationship between a pulsar's Dispersion Measure (DM) and its Signal-to-Noise Ratio (SNR), which is used to analyze and identify pulsar candidates. The "Pulsar" data-set contains information about various attributes of stars, such as the mean of the integrated profiles, standard deviation of the integrated profile, mean of the DM-SNR curve, standard deviation of the DM-SNR curve, and the target class.

In this paper, a comprehensive understanding of Support Vector Machine fundamentals are provided. The paper also undertakes a detailed exploration of the targeted problem and elucidates how the resulting insights are supported by a combination of visual representations and quantitative evaluation.

II. Support Vector Machines

In this section, we will describe the mathematical details of Support Vector Machines. Support Vector Machines is a supervised learning algorithm used for modelling the relationship between two types of variables, the target(dependent) and features(independent). The following is an overview of the jargons and mathematics utilised in the Support Vector Machines approach:

- *Hyperplane:* A hyperplane is a (n-1) dimension subspace where "n" is the dimension of the feature space. The hyperplane divides the feature space into different regions for each class.
- *Support Vectors:* The data points that are closest to the hyperplane and have the greatest influence on defining the margin are known as support vectors. Support vectors are essential to SVMs because they establish the hyperplane's orientation and location.
- *Margin:* The margin in an SVM refers to the separation between the hyperplane and the support vectors from each class. In SVMs, maximizing the margin is the primary objective.
- *Kernel Function:* SVM uses a kernel function to implicitly map the data into a higher-dimensional feature space where a linear hyperplane separating the data classes will be simpler to identify. The linear, polynomial, sigmoid, and radial basis function (RBF) kernels are a few examples of common kernel functions.
- *Regularization Parameter:* The regularization parameter is a crucial hyperparameter in Support Vector Machines (SVMs) that manages the trade-off between maximizing

the margin and reducing classification errors. It also plays an essential role in influencing the trade-off between model complexity and accuracy.

A. Assumptions in Support Vector Machines

Support Vector Machine is a powerful statistical technique, but its validity and reliability rests upon certain assumptions that must be met for accurate and meaningful results.

- *Scaled Features:* For the features to have an equal impact on the hyperplane, they must be standardized or normalized. By feature scaling, the optimization process is prevented from being dominated by a single feature.
- *Statistical Independence of Support Vectors:* The support vectors are assumed to be statistically independent by SVMs. Instead of using the complete data-set to make decisions, the Support Vector Machine (SVM) focuses only on the support vectors closest to the decision hyperplane.
- *Kernel Function Selection:* SVM uses kernel functions to map non-linearly separable data into a higher-dimensional space where it becomes linearly separable. The effective functioning of the SVM depends on selecting the correct kernel function.

B. Types of kernel functions

- *Linear:* The linear kernel is the most basic kernel function. It simply computes the dot product of two data points and is generally used only when the data is linearly separable.
- *Polynomial:* A more powerful kernel function is the polynomial kernel, which can be used to learn nonlinear relationships between data points.
- *Radial Basis:* Another powerful kernel function that can be used to learn nonlinear relationships between data points is the Radial Basis kernel or the Gaussian kernel. The Radial Basis kernel is more versatile than the polynomial kernel and is less prone to overfitting the data.
- *Sigmoid:* Neural networks are the source of the sigmoid kernel. It is similar to a neural network's activation function.

C. Algorithm

We will now explain the steps involved in modelling a Support Vector Machine:

- Standardize the data-set before feeding into the model, so that the process is not dominated by one feature.
- Select an appropriate kernel function based on the problem.
- The data will be transformed by the SVM using the kernel into a higher-dimension space.
- The SVM then finds the optimal hyperplane which maximizes the margin. It is accomplished by the following optimization problem:
  The data points are, $x_1, \ldots, x_n \in \mathbb{R}^p$ and the associated class labels are $y_1, \ldots, y_n \in \{-1, 1\}$. Let M be the margin that we want to maximize and K be the kernel function that we chose for this problem. Then,

  For $i = 1, \ldots, n$,

$$\underset{\beta_0, \alpha_0, \alpha_1, \ldots, \alpha_p, M}{\text{maximize}} M \tag{1}$$

$$y_i \left(\beta_0 + \sum_{i \in \mathcal{S}} \alpha_i K\langle x_j, x_i \rangle \right) \geq M \tag{2}$$

$$\text{subject to } \sum_{j=1}^{p} \alpha_j^2 = 1 \tag{3}$$

  Eqn. 2 ensures that each observation will be on the correct side of the hyperplane, given that M is positive. Eqn. 3 ensures that there exists one unique hyperplane for the chosen problem. Hence, this algorithm maximizes the margin M and obtains the separating hyperplane.
- Then, evaluation of the model should be done using the validation data-set.

D. Evaluation Metrics

We now define two quantities which can be used to evaluate the model.

- *F_1-score:* It is defined as the harmonic mean of precision and recall,

$$F_1 = \frac{2 * precision * recall}{precision + recall} \tag{4}$$

- *Accuracy:* It is the total number of correct predictions divided by total number of validation points,

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

- *ROC-AUC score:* The ROC-AUC (Receiver Operating Characteristic Area Under the Curve) is a graphical representation of a binary classification model's performance. It plots the True Positive Rate against the False Positive Rate across various probability thresholds for classifying positive and negative instances.

E. Advantages and Disadvantages of Support Vector Machines

- *Advantages:* Support Vector Machines have several advantages, including high-dimensional space effectiveness, robustness to outliers, and the ability to handle non-linearly separable data using kernel functions.
- *Disadvantages:* Support Vector Machines are inherently designed for binary classification. It also assumes that the data is linearly separable or can be transformed into a higher-dimensional space where it is separable linearly.

III. The Data

A data-set with information about various attributes of neutron stars, such as their mean of the integrated profiles, 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, and the target class are provided to us.

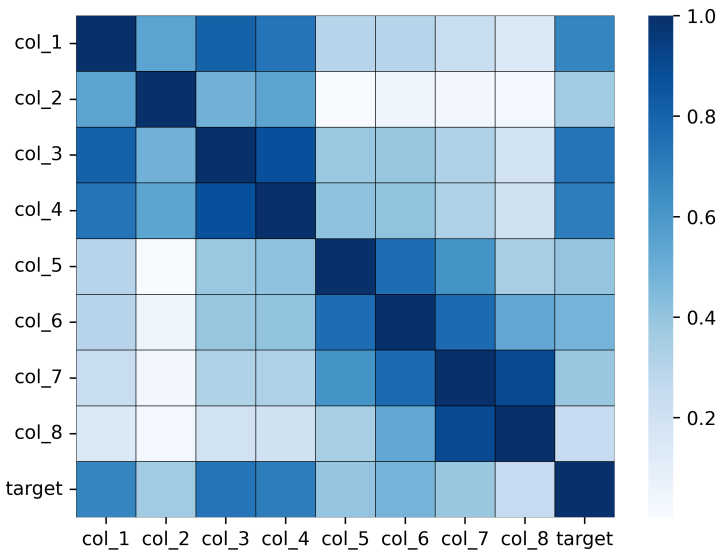

Fig. 1. Correlation heat-map within the variables

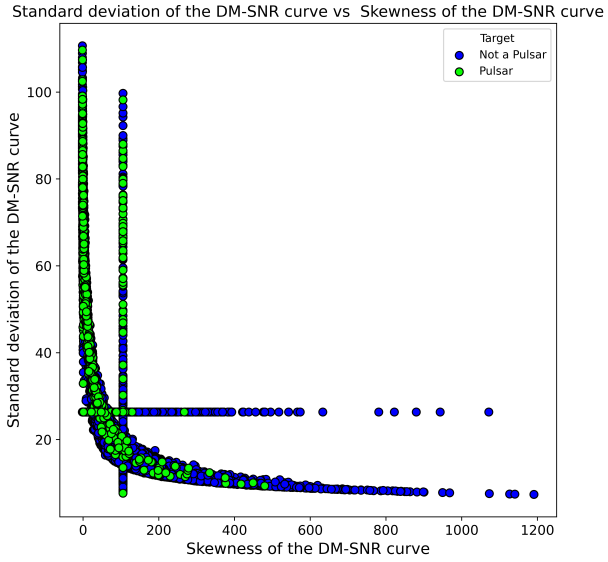

Fig. 3. Standard deviation of the DM-SNR curve vs Skewness of the DM-SNR curve

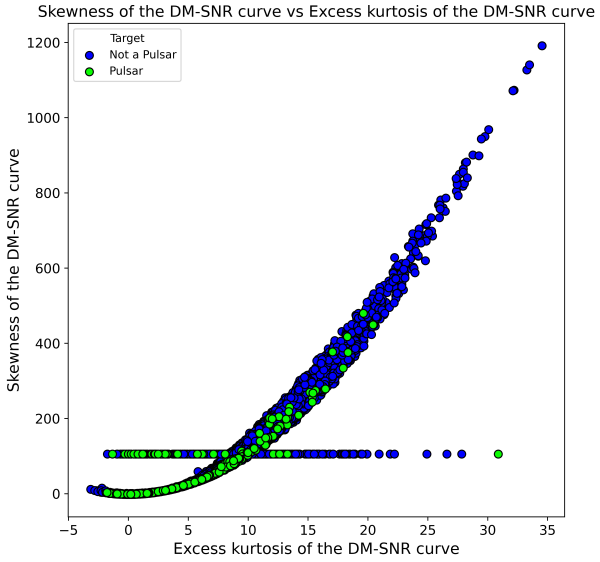

Fig. 2. Skewness of the DM-SNR curve vs Excess kurtosis of the DM-SNR curve

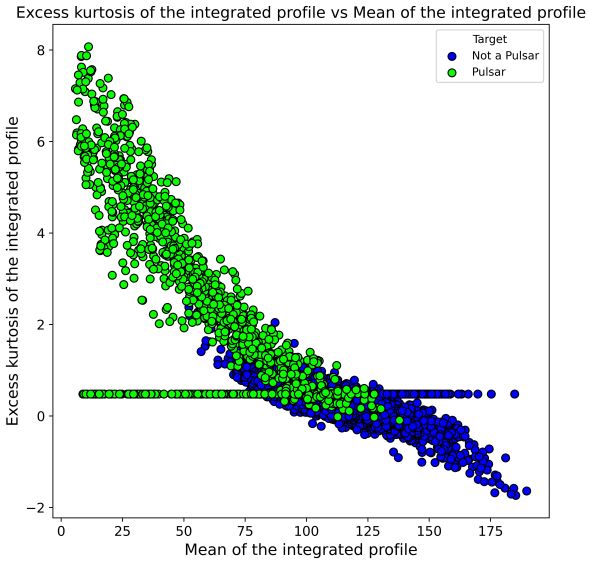

Fig. 4. Excess kurtosis of the integrated profile vs Mean of the integrated profile

Skewness, is the normalized third central moment about the mean. It characterizes the asymmetry of a data-set's distribution around its mean. A positive skewness means a heavy right-tail, negative skewness implies a heavy left-tail, and a meagre value of skewness indicates symmetric distribution.

Kurtosis, is the normalized fourth central moment about the mean. It's a statistical metric that characterizes how a distribution's tails are shaped in respect to normal distribution. A low kurtosis suggests light tails, and a high kurtosis suggests heavy tails.

The ultimate aim of the study is to classify a neutron star based on its attributes.

A. *Data description*

In this section, we describe the structure of the data-set. The train-set has 12,528 entries. It has been split into train(80%)-validation(20%) data-set. The test-set has 5370 entries.The columns from the data-set are listed below:

- col_1: *Mean of the integrated profile*
- col_2: *Standard deviation of the integrated profile*
- col_3: *Excess kurtosis of the integrated profile*
- col_4: *Skewness of the integrated profile*
- col_5: *Mean of the DM-SNR curve*
- col_6: *Standard deviation of the DM-SNR curve*
- col_7: *Excess kurtosis of the DM-SNR curve*

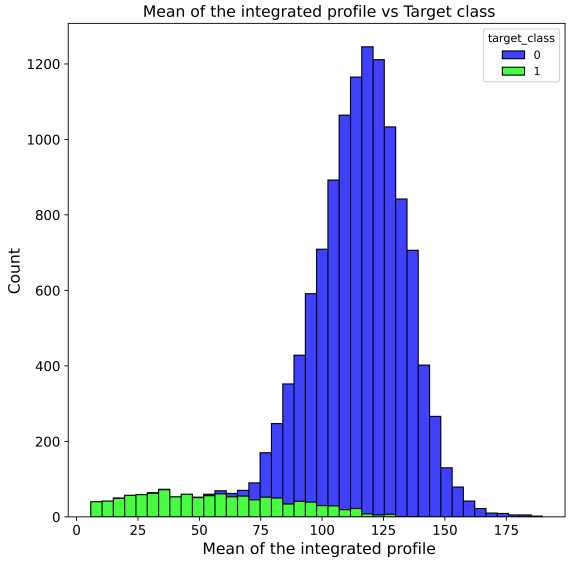

Fig. 5. Mean of the integrated profile vs Target class

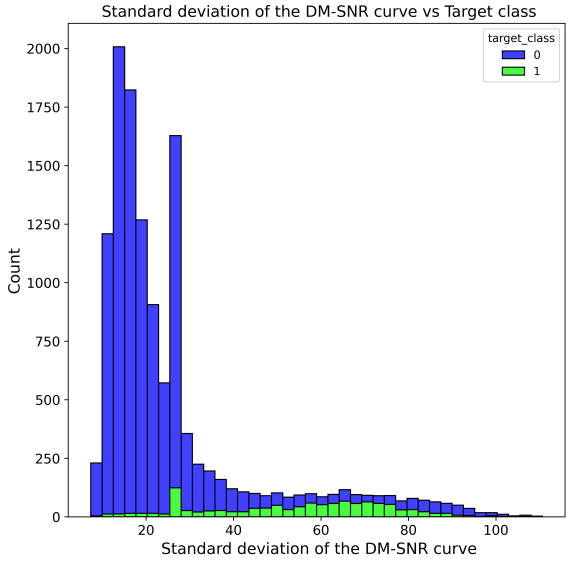

Fig. 6. Standard deviation of the DM-SNR curve vs Target class

- col_8: *Skewness of the DM-SNR curve*
- target: *target_class*

B. Data cleaning

While cleaning the data, we have to handle missing values. They were dealt in the following way:

- The data-set contained missing values only in the predictor variables. Precisely about 13.85% of the *Excess kurtosis of the integrated profile* data was missing, 9.40% of the *Standard deviation of the DM-SNR curve* data was missing, and 4.99% of the *Skewness of the DM-SNR curve* data was missing, which were imputed using respective mean, as all of the percentages are less that 15%.

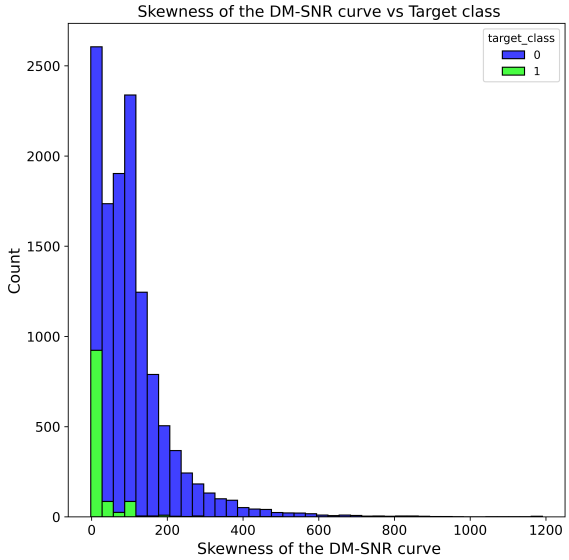

Fig. 7. Skewness of the DM-SNR curve vs Target class

- The highly correlated columns were removed as it affects the model performance. Then the data-set was standardized using the Gaussian standardization before feeding to the SVM model.

C. Data Visualization

1) Correlation plot: Fig. 1 shows the heat map of the correlation values between variables in the data-set. We can observe that a lot of attributes are correlated, hence they have to be removed.

From the correlation plot we can conclude that:

- The pair of variables with correlation greater than 0.75 are:
  - Excess kurtosis of the integrated profile & Mean of the integrated profile
  - Skewness of the integrated profile & Excess kurtosis of the integrated profile
  - Standard deviation of the DM-SNR curve & Mean of the DM-SNR curve
  - Excess kurtosis of the DM-SNR curve & Standard deviation of the DM-SNR curve
  - Skewness of the DM-SNR curve & Excess kurtosis of the DM-SNR curve
- The target variable is highly correlated with *Mean of the integrated profile*, *Excess kurtosis of the integrated profile*, and *Skewness of the integrated profile*.

2) Skewness of the DM-SNR curve and Excess kurtosis of the DM-SNR curve: From fig. 2, we can observe a high correlation of 0.90 between these attributes. We can also notice that pulsars are clustered near the origin, indicating low skewness and kurtosis in the DM-SNR curve.

3) Standard deviation of the DM-SNR curve and Skewness of the DM-SNR curve: We can observe from fig. 3 that, there is a correlation of 0.53 between these attributes. We can

also notice that pulsars are distributed all across the standard deviation of the DM-SNR curve, hence this attribute does not have much influence on the target class.

4) Excess kurtosis of the integrated profile and Mean of the integrated profile: From fig. 4, we can observe a high correlation of 0.81 between these attributes. We can also notice that pulsars are clustered near the lower values of the mean of integrated profile.

5) Mean of the integrated profile and Target class: We can observe from fig. 5 that, non-pulsars have a higher value of mean of the integrated profile and pulsars are distributed near the lower values of the mean of integrated profile, denoting an even distribution of star mass about the center, in pulsars.

6) Standard deviation of the DM-SNR curve and Target class: We can notice from fig. 6 that, non-pulsars have a low standard deviation value and pulsars have a higher standard deviation indicating that, pulsars have a heavy tail.

7) Skewness of the DM-SNR curve and Target class: We can notice from fig. 7 that, non-pulsars are distributed across all skewness values but, pulsars have a low skewness indicating symmetric distribution about the center of the star.

IV. The Problem

A. Outline

We have done the analysing, cleaning and visualizing the data. A train-set and a validation-set are created from the data-set. The validation-set is used to make predictions, after the Support Vector Machine model has been trained using the train-set. The validation-set predictions are used to calculate three different metrics to assess model performance.

B. Leave One Out Cross-Validation

To obtain the hyper parameter C, which defines the regularization value of the soft-margin classifier, we use a special cross-validation called Leave One Out Cross-Validation(LOOCV). It is a cross-validation technique used to find the right hyper-parameters by training the model on subsets of the data-set and testing on the unseen data.

LOOCV will be explained below in the context of our case, where we have to tune the regularization value of the soft-margin classifier(C).

- Firstly, we take 5 values of the hyper-parameter $C = [0.1, 0.5, 1, 5, 10]$.
- Then we split the data-set into 5 subsets $\{(X_1, y_1), (X_2, y_2), (X_3, y_3), (X_4, y_4), (X_5, y_5)\}$.
- Now for each C value, we randomly leave out one of these subsets and train the SVM model(with the current iterating value of C) on the remaining 4 subsets of data, and then evaluate the accuracy score on the left out subset.
- The results of LOOCV are reported in TABLE I.
- Hence we choose $C = 10$ as our soft-margin regularization value.

TABLE I
LOOCV results

C	R^2-score
0.1	0.973
0.5	0.976
1	0.977
5	0.977
10	0.978
15	0.977
20	0.977
25	0.977

C. Support Vector Machine modelling

Before fitting the Support Vector Machine to the training data by using *scikit-learn* package in Python, we need to specify the *kernel* parameter of the SVM, which denotes the kernel function used in the model. In order to accomplish this, we try all the possible kernel options. We observe that ***RBF-kernel*** gives the maximum accuracy.

We now instantiate a Support Vector Machine model, with parameters (*kernel='rbf'*, $C = 10$), for predicting the target class. We use F_1*-score*, *Accuracy* and *ROC-AUC* scores as our evaluation criteria on the validation data-set. After testing on the Validation Split, the model obtains a F_1*-score* of 0.99 an *Accuracy* score of 0.98, and *ROC-AUC* score of 0.95.

D. Confusion matrix

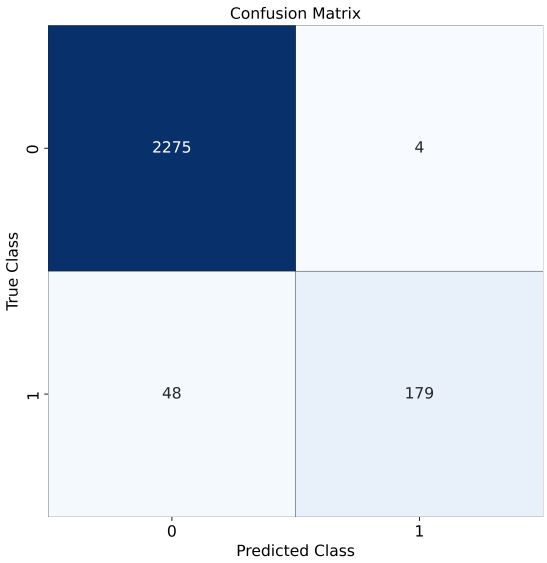

Fig. 8. Confusion Matrix

Fig. 8 represents the confusion matrix generated and gives the count to all the possible outcomes in the validation set(20% of the data). We can notice that the True positives and False negatives are very high compared to the True negatives and False positives.

E. ROC-AUC score

We can notice that the area under the curve (Fig. 9) is close to 1, and the curve tends towards the LHS top corner, which

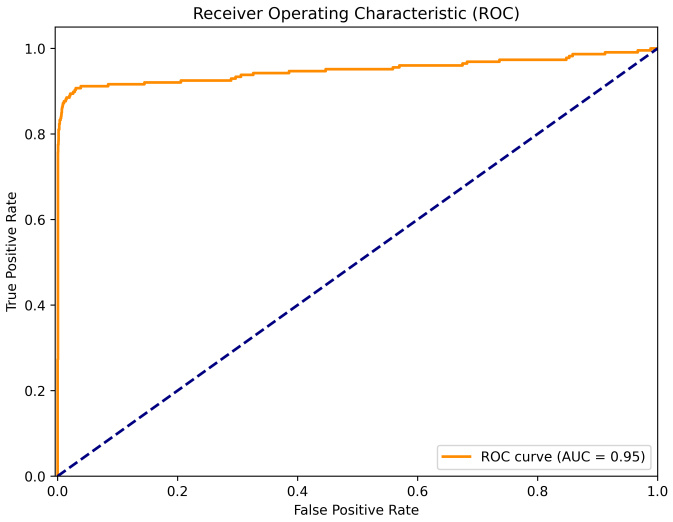

Fig. 9. ROC-AUC curve

implies it has a high true positive rate and low false positive rate, hence more clearer differentiation between classes.

V. Conclusion

The conclusions that can be drawn from all the above analysis are:

- The Radial Basis Function (RBF) kernel was found to be the most effective for this data-set.
- The high value of evaluation metrics denotes that the Support Vector Machine (SVM) model effectively classified pulsars from non-pulsars.
- Even though the data-set had class imbalance, SVM handled it in a robust way.
- We can conclude that the measures of central tendencies of a neutron star are important in classifying pulsars from non-pulsars.
- We have also predicted the target class for the test-set provided using the best SVM model.

Future research could use soft-margin classifier, that allows individual data-points to be on the wrong side and hence, creating a more robust model. We can also use multiple kernels that can enhance the model's ability to capture complex patterns in the data. We could also combine SVMs with other machine learning techniques to create powerful ensemble models. We have observed that, there is class imbalance in the target class column. This may create inherent bias in the model, hence we can adopt class-balancing techniques to overcome this bias.

References

[1] "Support vector machine." Wikipedia, Wikimedia Foundation, 4 Nov. 2023, *https://en.wikipedia.org/wiki/Support_vector_machine* (accessed 8 Nov. 2023)

[2] Imagine the Universe! (n.d.). *https://imagine.gsfc.nasa.gov/science/objects/neutron_stars1.html* (accessed 8 Nov. 2023)

[3] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An Introduction to Statistical Learning : with Applications in R*. New York :Springer, 2013.