SPARTIFICIAL

PREDICTING PULSAR STARS USING CLASSICAL MACHINE LEARNING

Internship Id: ads-1209-gr1

Team Members:

- Amoy Ashesh
- Aquil Ahmed
- Viraj Parmaj
- Anisha Yadav
- Aditi Sharma

Instructor:

- Dr. Kavitha C.

INDEX

1.	ABSTRACT	3
2.	INTRODUCTION	4
3.	MOTIVATION	6
4.	LITERATURE SURVEY & REVIEW	8
5.	METHODOLOGY	12
6.	PERFORMANCE REPORT & RESULTS	20
7.	CONCLUSION & FUTURE SCOPE	26
8.	REFERENCES	28
9.	PROJECT TIMELINE	29

ABSTRACT

Pulsar stars are fascinating astronomical objects that emit regular pulses of electromagnetic radiation. Identifying pulsars is essential to understanding various properties of the universe, such as the nature of gravity, the properties of matter in extreme conditions, and the evolution of stars. However, distinguishing pulsars from other objects can be challenging due to the low signal-to-noise ratio of their emissions.

Classical machine learning algorithms have recently become increasingly popular in astronomical research for classifying and predicting various astronomical objects. The researchers use several machine learning algorithms, including decision trees, random forest, and support vector machine, to determine whether a candidate object is a pulsar or not. In this study, the dataset included over 17,000 observations of pulsar candidates, each with eight features related to physical properties, such as the mean and standard deviation of the integrated pulse profile and the mean and standard deviation of the DM-SNR curve.

The results from this study showed that the random forest algorithm achieved the highest accuracy, with a rate of 98.04%. The random forest algorithm combines multiple decision trees to improve the model's accuracy and minimise the risk of overfitting. The logistic regression and support vector machine also performed well, achieving accuracies of 97.91% and 97.93%, respectively.

This study highlights the usefulness of classical machine learning algorithms in identifying pulsar stars and emphasises the importance of optimising feature selection and algorithm selection to improve accuracy. Accurately identifying pulsar stars is crucial for advancing our understanding of the universe's properties and evolution, making this study relevant for future astronomical research.

INTRODUCTION

While smaller than a huge city and spherical and compact-like, pulsars have more mass than the sun. Scientists use pulsars to investigate the most extreme forms of matter, look for extrasolar planets, and measure cosmic distances. Pulsars may also aid in the discovery of gravitational waves, which may lead to the detection of powerful cosmic occurrences like mergers of extremely large black holes. Pulsars, which were first identified in 1967, are intriguing constituents of the cosmos.

Pulsars emit two constant, directional narrow beams of light. Pulsars appear to flicker even if their light is stable because they also spin. The same factor accounts for why a lighthouse appears to blink to a sailor travelling by sea: The light beam may sweep over the Planet as the pulsar rotates, swing out of sight, and then swing back around. The light appears to come and go, creating the impression that the pulsar is blinking to an observer on the earth, which is why pulsars are often termed Cosmic Lighthouse. The light beam of a pulsar often is not aligned with the pulsar's axis of rotation, which is why it spins around like a lighthouse beam.

To speed up analysis, pulsar candidates are now automatically labelled using machine learning methods. The ubiquitous use of classification systems that approach the candidate data sets as binary classification issues. We have used various machine-learning classification techniques to show how pulsar star predictions are made.

To get the required results, our model will employ supervised learning. In general, when a system is given input and output variables to learn how they are mapped together or connected, this process is known as supervised learning. The objective is to create a mapping function that is accurate enough for the algorithm to anticipate the output from new input. Every time the algorithm generates a prediction, it is corrected or provided feedback, and this process is repeated until the algorithm performs at an acceptable level.

Nevertheless, radio repeat impedance and fuss can eventually produce signals that match those of pulsars; therefore, it is fascinating to consider a method to distinguish between pulsars and radio repeat blockage or noise.

Five models are being used in this article to determine the existence of pulsar stars. The integrated profile's mean, standard deviation, excess kurtosis, and skewness, as well as the DM-SNR curve's mean, standard deviation, skewness, and excess kurtosis, are included in the feature set.

Our issue necessitates the use of classification methods. In a nutshell, classification either creates a model based on the training set and the values (class labels) of the classifying variables and uses it to classify incoming data, or it predicts categorical class labels. Many categorisation models exist.

The logistic regression, decision tree classifier, random forest classifier, K-Nearest-Neighbours, and Support Vector Machine are the five ML models we have included in our code to provide a predicted label that may be either 1 or 0 (where 1 denotes a pulsar star and 0 denotes a non-pulsar star).

MOTIVATION

Understanding pulsar stars is crucial for comprehending the universe's many characteristics, such as gravity, matter under severe circumstances, and star development. Identification of pulsars, a rare astronomical phenomenon that emits electromagnetic radiation in regular pulses, can shed light on the behaviour of matter under conditions that cannot be seen in a laboratory. Pulsars can also be effective research instruments for figuring out the nature of gravity and putting Einstein's theory of relativity to the test. They can also reveal important details about the characteristics of neutron stars, some of the universe's densest objects. Scientists can better understand the underlying nature of the world and the physical rules that govern it by studying pulsars.

Here are some of the key reasons why we study pulsar stars in more detail:

1. Study of properties of the universe

The interesting celestial phenomenon known as a pulsar produces electromagnetic radiation pulses on a regular basis. Scientists can better understand the underlying nature of the world and the physical rules that govern it by studying pulsars. For instance, pulsars can aid in our understanding of gravity, a fundamental force of nature. Because of their great density and powerful gravitational fields, pulsars are extraordinarily dense objects whose study might shed light on how gravity behaves under challenging circumstances. Pulsars can also aid in our understanding of the creation and development of galaxies, stars, and the cosmos in general.

2. Investigating Material Behaviour under Severe Environments

Pulsars are very dense objects composed of strange substances like quarks and gluons. By examining pulsars, scientists can better understand how matter behaves under severe circumstances that are not visible in a laboratory. This has potential implications in areas like material science and engineering, where it is crucial to comprehend the characteristics of matter under severe circumstances.

3. Testing the Physical Laws

The rules of physics can also be tested using pulsars. For instance, the presence of gravitational waves, which are rippling effects of the motion of enormous objects in space-time, is predicted by Einstein's theory of general relativity. Pulsars can aid in the detection of gravitational waves, which can provide essential details about the origin and development of the cosmos. Scientists may test the predictions of general relativity and other physics theories by analysing the effects of gravitational waves on pulsars.

4. Development of Technology

New technology may be created as a result of pulsar research. For instance, the methods used to find and study pulsars can be modified for use in other contexts, such as hunting for exoplanets or detecting gravitational waves. Moreover, pulsar research can provide ideas for novel materials or propulsion systems.

While improving our knowledge of the cosmos and its characteristics is the primary goal of pulsar star research, it may also be utilised to solve issues in the real world. For instance, pulsar research can further our knowledge of gravitational waves, which are disturbances in space-time's fabric brought on by the movement of enormous objects.

Finding gravitational waves can help us understand the origins and development of the cosmos. It can also be used in physics, cosmology, and other scientific disciplines. Moreover, research on pulsars can assist in improving models of how matter behaves under severe circumstances, which has implications for engineering and material science. Finally, knowledge of the characteristics of neutron stars, which are connected to pulsars, can have effects on disciplines like nuclear physics and condensed matter physics. In conclusion, while the primary goal of pulsar research is to further our understanding of the cosmos, it may also be employed in various real-world settings.

LITERATURE SURVEY & REVIEW

Introduction:

Pulsars are highly magnetised neutron stars that rotate quickly and generate electromagnetic radiation in the form of beams. They have been the focus of a considerable investigation by astronomers and astrophysicists since their discovery in 1967, yielding significant findings and insights into the nature of the cosmos. Recent research has proven that machine learning is a powerful technique for pulsar prediction from astronomical data and has shown encouraging results. This literature review provides an overview of the state-of-the-art pulsar prediction using machine learning.
Here are a few more thorough descriptions of the pulsar research:

1. Pulsar Timing

Time of pulsars: Pulsars are extremely reliable clocks that periodically release electromagnetic pulses at highly exact intervals. Astronomers may learn about the pulsar's rotational speed, magnetic field, and the characteristics of the interstellar medium by studying the arrival timings of these pulses. Because of this research, several pulsars have been found, including the millisecond pulsars, which are among the universe's most stable clocks. Gravitational waves, which are ripples in space-time created by the acceleration of massive objects, have also been discovered using pulsar timing.

2. Gravitational Waves

Gravitational waves, which are ripples in space-time created by the acceleration of enormous objects, have been discovered using the exact timing of pulsars. Using a

network of pulsars scattered throughout the sky, pulsar timing arrays have been utilised to find gravitational waves in the nano-hertz frequency range. Pulsar timing arrays enabled the Laser Interferometer Gravitational-Wave Observatory (LIGO) to make the first gravitational wave detection in 2015.

3. Neutron Star Structure

As neutron stars are very dense objects created from the remains of supernova explosions, pulsars offer a rare glimpse into their structure. Astronomers can learn more about the behaviour of matter in extreme situations by examining the pulsar's characteristics. For instance, the study of pulsar rotational anomalies has revealed details about the internal structure of neutron stars, including the potential for a superfluid core. The equation of state of matter at high densities, which has significant ramifications for nuclear physics and astrophysics, has also been studied using pulsars.

4. Magnetospheres

Strong magnetic fields produced by pulsars result in complex, dynamic magnetospheres. Astronomers can learn more about the physics of plasmas and the behaviour of magnetic fields in astrophysical settings by investigating the characteristics of these magnetospheres. As an illustration, research on pulsar magnetospheres has revealed phenomena like magnetospheric emission and magnetar flares, which are considered to be brought on by the fast loss of magnetic energy in the neutron star's magnetosphere.

5. Pulsar Planets

The discovery of planets around several pulsars in recent years has shed light on the emergence and development of planetary systems around neutron stars. Typically, pulsar planets are found by observing the gravitational modulation of the pulsar's

radio signal. These discoveries raised the genesis, development, and possible habitability of planets circling pulsars, raising new issues.

Recent years have seen several fresh discoveries and understandings regarding pulsars. Here are a few illustrations:

(a) Fast Radio Bursts (FRBs):

Fast Radio Bursts are millisecond-long, intense radio emission bursts that seem to originate outside our galaxy. Several FRBs have recently been discovered from magnetars, strongly magnetic neutron stars connected to pulsars. The precise process of magnetars creating the bursts is unclear, but these observations imply that they may be responsible for at least some FRBs.

(b) Pulsar Glitches:

A neutron star's crust is expected to interact with its superfluid innards to generate pulsar glitches, which are abrupt variations in a pulsar's rotational rate. One of the most extensively researched pulsars, the Crab Pulsar, has had many problems recently. These observations have provided a fresh understanding of the pulsar's internal structure and the behaviour of matter in severe circumstances.

(c) Polarisation

Important details regarding the structure of the pulsar's magnetic field and the characteristics of the emission process may be learned from the polarisation of the pulsar radiation. Several pulsars have recently been seen to display complicated polarisation patterns, which are assumed to be a result of the presence of powerful magnetic fields and intricate plasma environments.

(d) Multi-Messenger Astronomy

Several telescopes, including radio, X-ray, and gamma-ray, have been used to observe pulsars. Astronomers can better understand the behaviour of pulsars and their environs by integrating data from these many instruments. For instance, the Fermi Gamma-ray Space Telescope's investigations of pulsars have revealed pulsars with extraordinarily high-energy gamma-ray emission, which has revealed new details about the processes that accelerate high-energy particles in the pulsar magnetosphere.

Background:

Pulsars are highly compact neutron stars that rotate quickly and generate powerful radiation beams in a distinctive pattern. Radio telescopes can detect the radio waves that pulsars emit, but noise and interference in the data and the resemblance of pulsar signals to those of other astronomical events can make pulsar detection challenging. Machine learning algorithms can detect pulsars with great accuracy thanks to their ability to recognise patterns in data and forecast outcomes based on those patterns.

Approach:

The subject of pulsar prediction has been tackled in recent years using a range of machine-learning approaches. Deep learning, unsupervised learning, and supervised learning are some of these methods.

METHODOLOGY

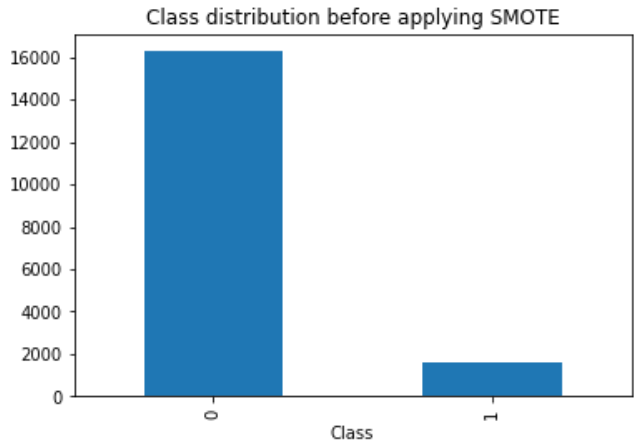
1. Balancing the dataset:

In data analysis and machine learning, having an imbalanced dataset can significantly impact the accuracy of the resulting predictions. In such cases, Synthetic Minority Oversampling Technique (SMOTE) is a commonly used method to balance the dataset.

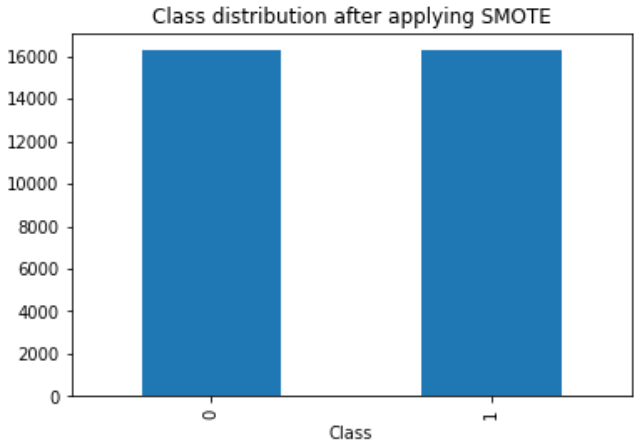
SMOTE is a technique that generates synthetic data points for the minority class to balance the distribution of the classes in the dataset. This technique creates new observations for the minority class by using interpolation methods to create "synthetic" samples that are similar to the existing minority class observations. This process continues until the minority class has a representation similar to that of the majority class.

By using SMOTE to balance the dataset, the resulting distribution of the classes is more even, which allows for more accurate predictions by machine learning models. This process can mitigate the issue of imbalanced classes and can lead to better results when working with imbalanced datasets.

Overall, SMOTE is a powerful technique for improving the accuracy of machine learning models when working with imbalanced datasets, and it is frequently used in data analysis and machine learning projects.

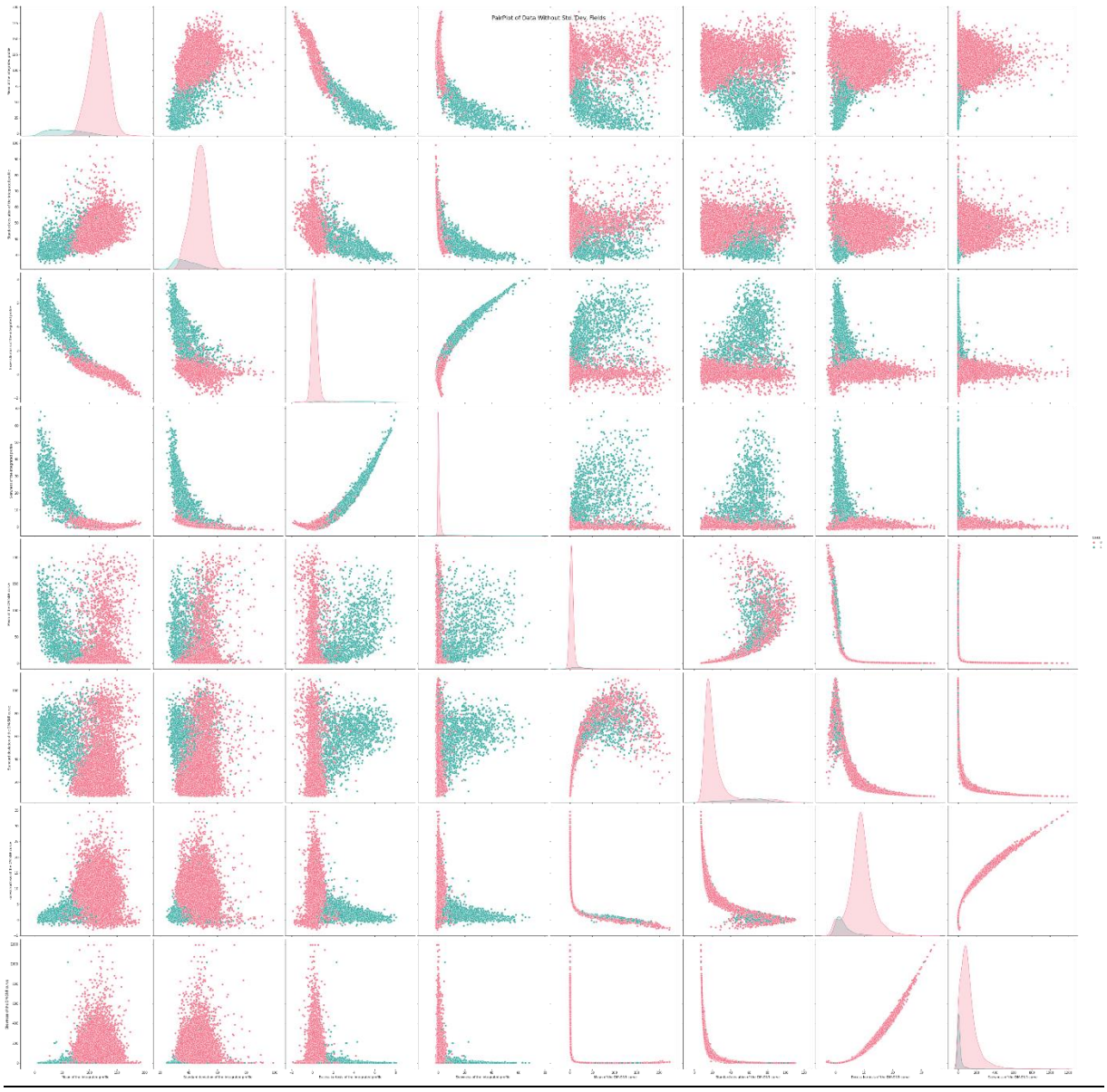


Fig[1]: Before applying SMOTE



Fig[2]: After applying SMOTE

2. Plotting the distribution of variables and finding the Correlation between the variables:



Figure[3.]: Visualising the variables in the dataset

In machine learning, the term "correlation" refers to the connection between two variables, where a change in one variable is related to a change in the other. The direction and strength of the association between two continuous variables are specifically mentioned. If two features are highly correlated, one of them can be

removed since they provide similar information. Correlation is also used to identify data patterns and build predictive models.

Correlation can be positive or negative, depending on the direction of the relationship between the variables. Positive correlation occurs when an increase in one variable is associated with an increase in the other variable, while negative correlation occurs when an increase in one variable is associated with a decrease in the other variable. The strength of the correlation can be measured by the correlation coefficient, which ranges from -1 to 1. A correlation coefficient of 0 indicates no correlation between the variables, while a correlation coefficient of -1 or 1 indicates a perfect negative or positive correlation, respectively.

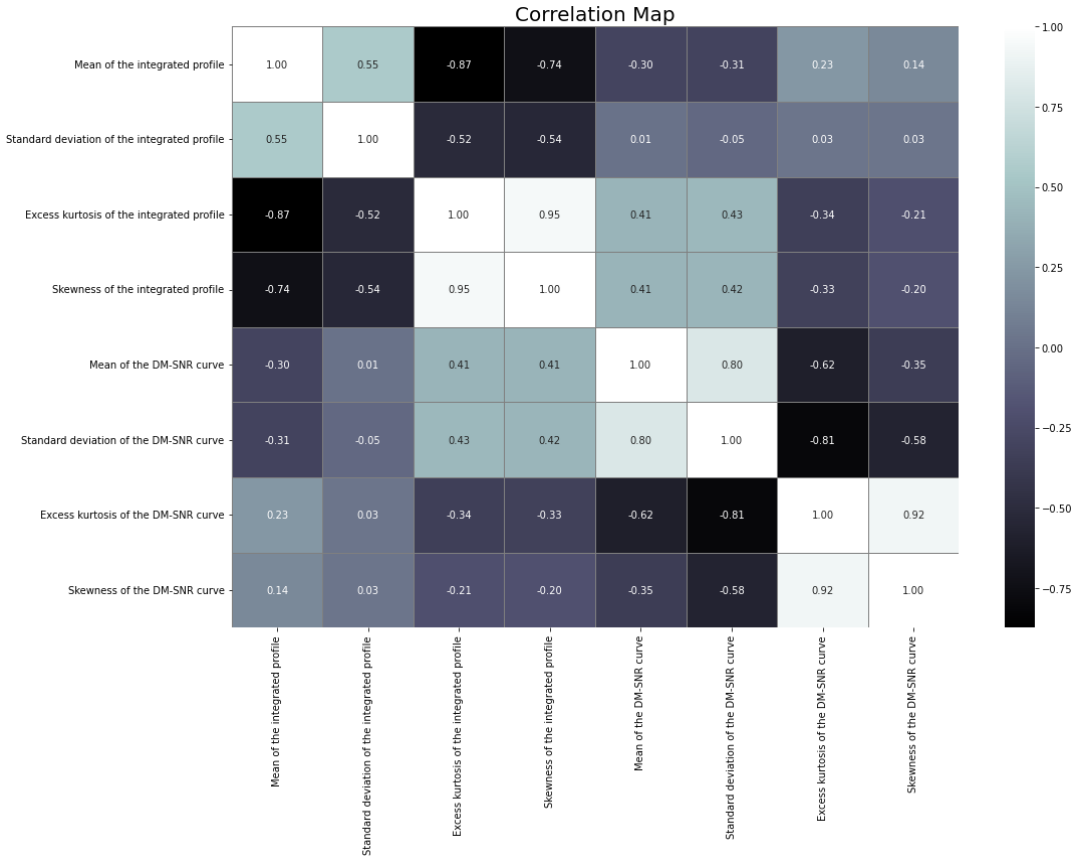


Figure [4.]: Correlation Map between all the variables in the dataset

3. Training the ML models:

The models are fitted onto a train dataset, and the dataset used for prediction is called the test dataset. The dataset was split into test and train cases in the ratio of 20: 80. The various Machine Learning models used are:

1. Logistic Regression:

   In machine learning, the Supervised Learning subcategory includes the commonly used algorithm of logistic regression. Its primary purpose is to predict the outcome of a dependent variable that belongs to a category based on a set of independent variables. This means that the output must be categorical or discrete in nature, such as Yes or No, 0 or 1, or true or false. However, instead of providing an exact value of 0 or 1, logistic regression gives probabilistic values that range between 0 and 1.

2. Decision Tree Classifier:

   The Decision Tree is a machine learning algorithm that falls under the category of supervised learning. It is used to solve classification and regression problems but primarily for classification tasks. A Decision Tree is a type of classifier that has a tree-like structure. The tree's internal nodes represent a dataset's features, while the branches represent the decision-making processes, and each leaf node represents the outcome or result. Decision Nodes and Leaf Nodes are the two types of nodes in a decision tree. Decision Nodes are responsible for making decisions and have multiple branches, while Leaf Nodes represent the final decision or output and have no other branches. The Decision Tree makes its decisions or tests based on the characteristics or properties of the given dataset.

3. Random Forest Classifier:

   Random Forest Classifier is a machine learning algorithm that belongs to the supervised learning technique. It is often used for classification tasks where the output variable is categorical or discrete. Random Forest Classifier is based on ensemble learning, combining multiple

decision trees to make more accurate predictions. The algorithm creates a forest of decision trees, each using a random subset of the features and data points.

During the training process, the algorithm randomly selects a subset of features and a subset of data points and creates a decision tree. This process is repeated several times to create multiple decision trees. The algorithm predicts by aggregating each tree's predictions and choosing the class that receives the most votes. This approach helps to improve the accuracy and robustness of the model, as it reduces the impact of individual trees that may be overfitting the data. Random Forest Classifier is often used for applications such as image classification, text classification, and bioinformatics.

4. K-Nearest Neighbours:

K-Nearest Neighbors (KNN) is a popular machine learning algorithm for classification and regression tasks. It falls under the category of supervised learning, meaning that it requires labelled data to train the model. In KNN, the algorithm classifies new data points based on their proximity to the k-nearest data points in the training dataset. The user determines the value of k and represents the number of nearest neighbours to consider. KNN is a non-parametric algorithm, meaning it makes no assumptions about the data distribution. It is also easy to understand and implement, making it a popular choice for many classification and regression tasks. However, its performance can be affected by choice of k, and it can be computationally expensive for large datasets.

5. Support Vector Machine:

Support Vector Machine (SVM) is a popular machine learning algorithm used for classification, regression, and outlier detection tasks. It is a supervised learning algorithm, meaning that it requires labelled data to train the model. In SVM, the algorithm constructs a hyperplane in a high-dimensional space that can be used to separate the different classes in the data. The goal is to find the hyperplane that maximises

the margin, which is the distance between the hyperplane and the nearest data points of each class.

Using the kernel trick technique, SVM can handle both linear and non-linearly separable data. The kernel trick transforms the input data into a higher-dimensional space that can be linearly separable. SVM is particularly useful when dealing with high-dimensional data, such as image or text classification tasks. It is also effective in handling datasets with a small number of samples, as it is less prone to overfitting compared to other algorithms. However, SVM can be computationally expensive for large datasets and can be sensitive to the choice of the kernel function and other hyperparameters. Nonetheless, with careful tuning of the parameters, SVM can be a powerful tool for solving many classification, regression, and outlier detection problems.

4. Evaluating the models:

The accuracy, recall, F1 measure, k-fold cross-validation score was evaluated on each ML algorithm. Confusion Matrices for all the algorithms were also plotted.

Accuracy is a measure of how often the model correctly predicts the outcome of a task. It is calculated by dividing the number of correct predictions by the total number of predictions. Accuracy can be a helpful metric when the classes in the data are well-balanced.

Recall is a measure of how well the model identifies positive instances. It is calculated by dividing the number of true positives by the sum of true positives and false negatives. Recall is a valuable metric for correctly identifying all positive instances, such as in medical diagnosis.

F1 measure is a combination of precision and recall, which provides a balance between these two metrics. It is calculated as the harmonic mean of precision and

recall and considers both false positives and false negatives. F1 measure is often used in binary classification problems when the data is imbalanced.

K-fold cross-validation is a technique used to evaluate the performance of machine learning models on a dataset. It involves dividing the dataset into k equal parts, training the model on k-1 parts, and testing it on the remaining part. This process is repeated k times, each part serving as the testing set once. The average performance of the model across all k iterations is used as the final evaluation metric.

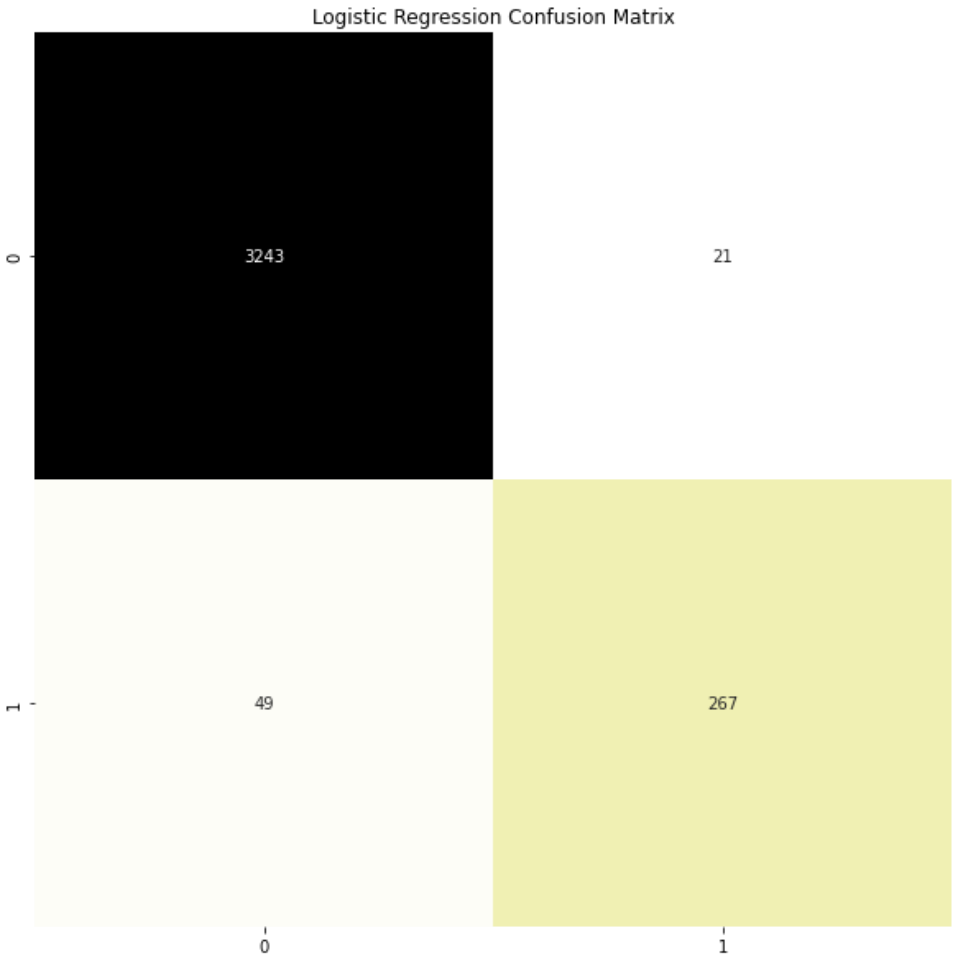
A confusion matrix is a table often used to evaluate the performance of a classification model. It is a matrix that summarises a model's output's predicted and actual classifications, providing a more detailed view of its performance than just a single accuracy score.
A confusion matrix has four main components: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). These components are defined as follows:

- True positives (TP): The number of instances where the model correctly predicted the positive class.
- False positives (FP): The number of instances where the model predicted the positive class, but the actual class was negative.
- True negatives (TN): The number of instances where the model correctly predicted the negative class.
- False negatives (FN): The number of instances where the model predicted the negative class, but the actual class was positive.

PERFORMANCE REPORT & RESULTS

1. Logistic Regression:

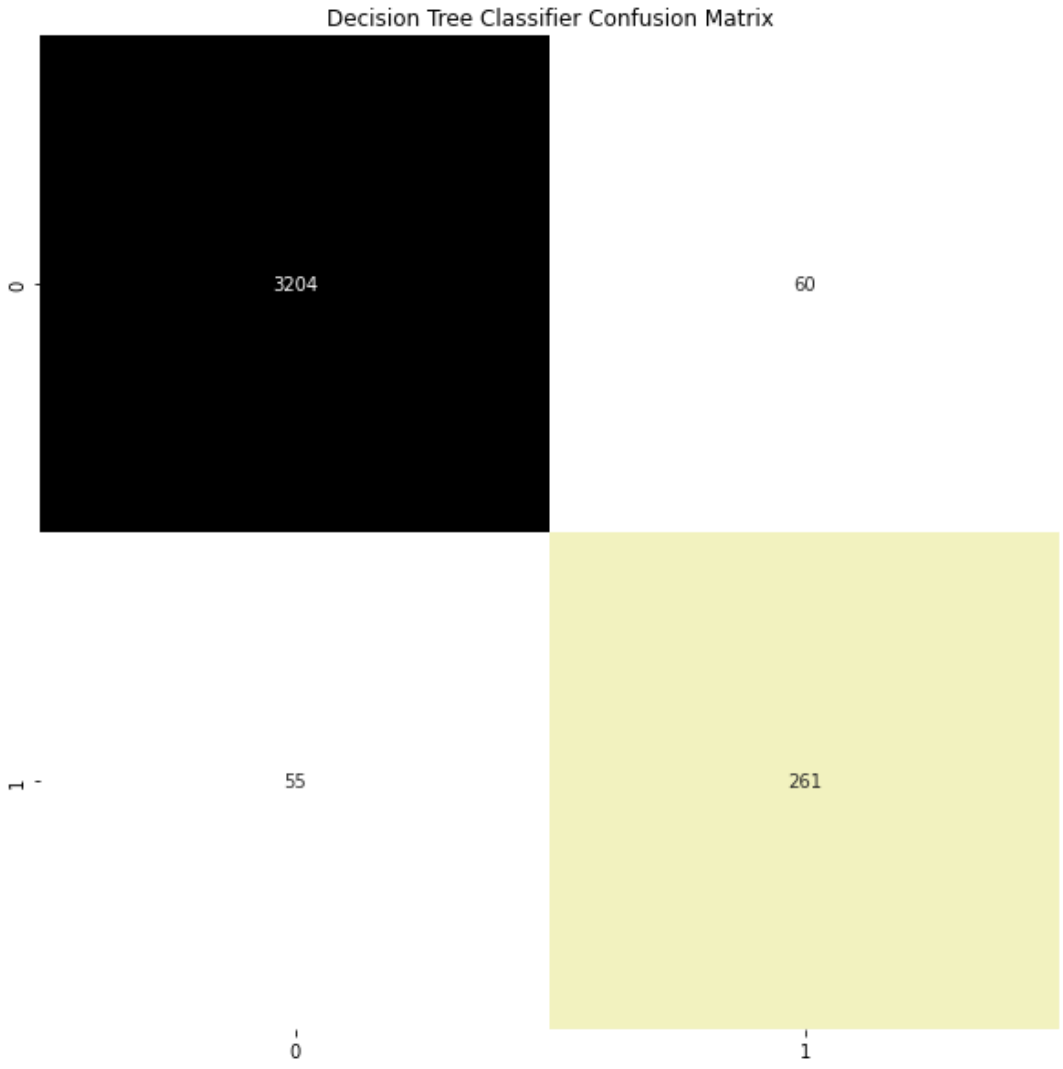
Accuracy: 0.9804469273743017
Recall: 0.8449367088607594
F1 measure: 0.8841059602649006
K-fold Cross Validation Score: 97.91% (0.23%)



Figure[5.]: Logistic Regression Confusion Matrix

2. Decision Tree Classifier:

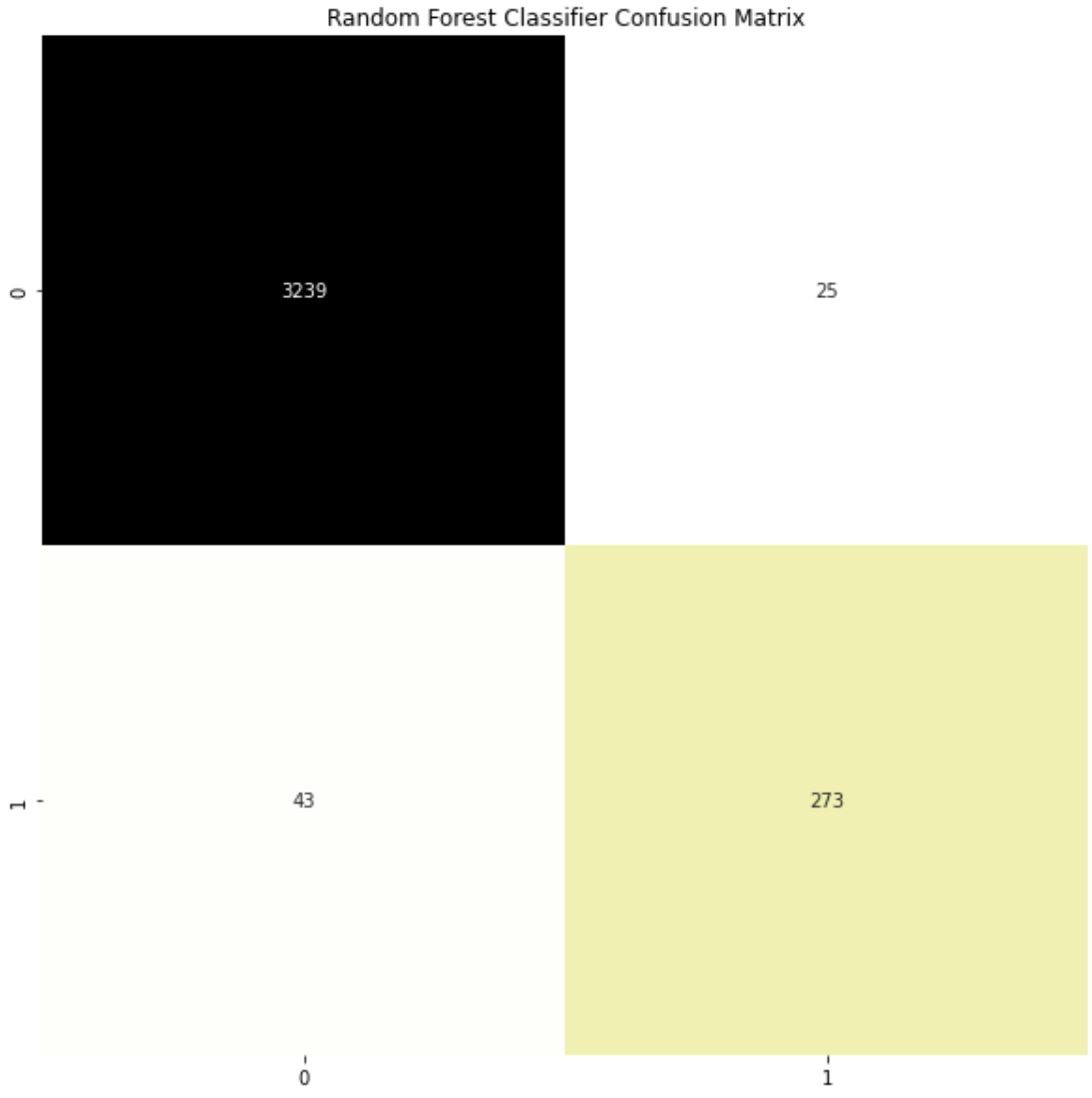
Accuracy: 0.9678770949720671
Recall: 0.8259493670886076
F1 measure: 0.8194662480376766
K-fold Cross Validation Score: 96.64% (0.35%)



Figure[6.]: Decision Tree Classifier Confusion Matrix

3. Random Forest Classifier:

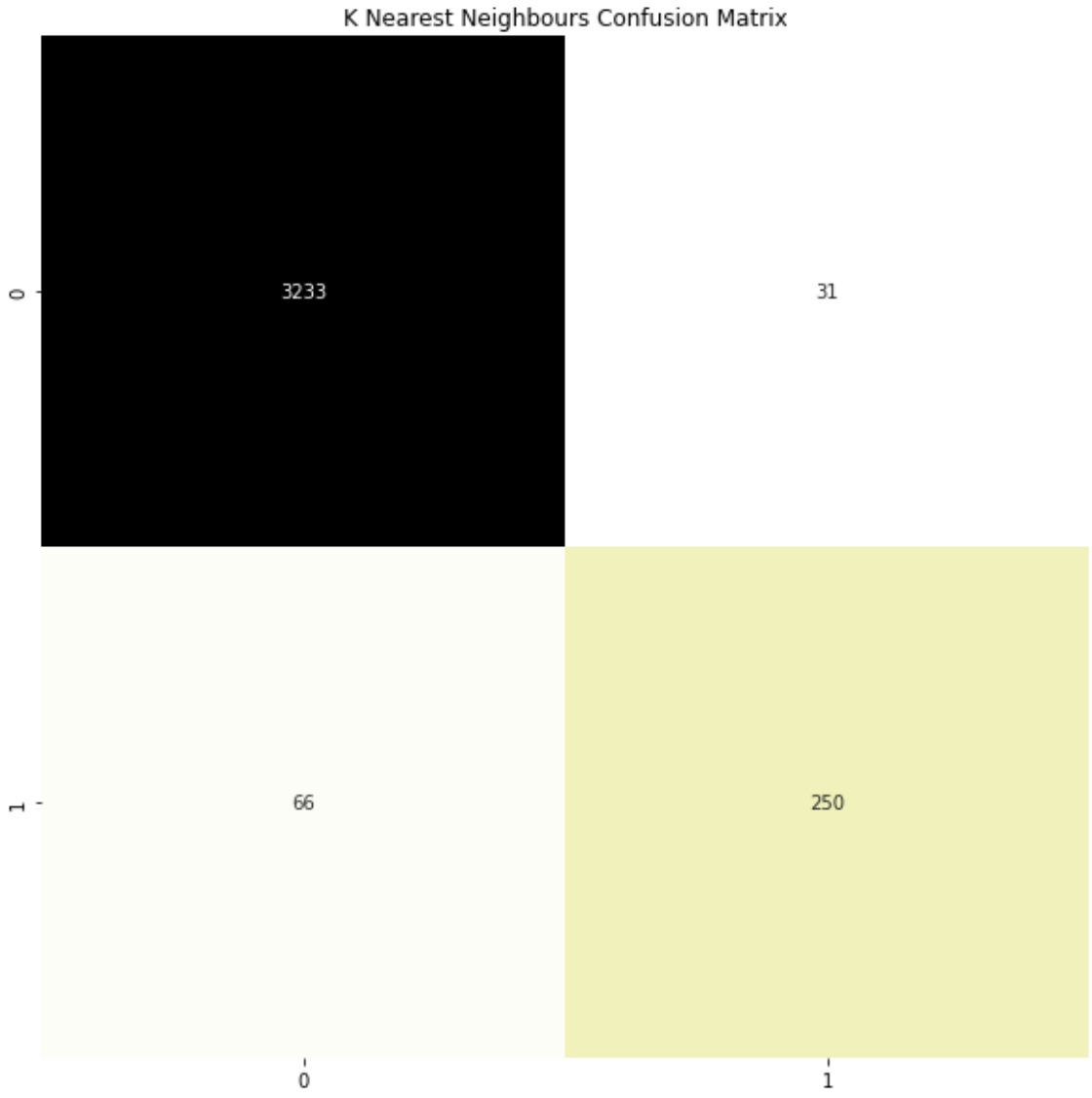
Accuracy: 0.9810055865921787
Recall: 0.8639240506329114
F1 measure: 0.8892508143322475
K-fold Cross Validation Score: 98.04% (0.34%)



Figure[7.]: Random Forest Classifier Confusion Matrix

4. K-Nearest Neighbours:

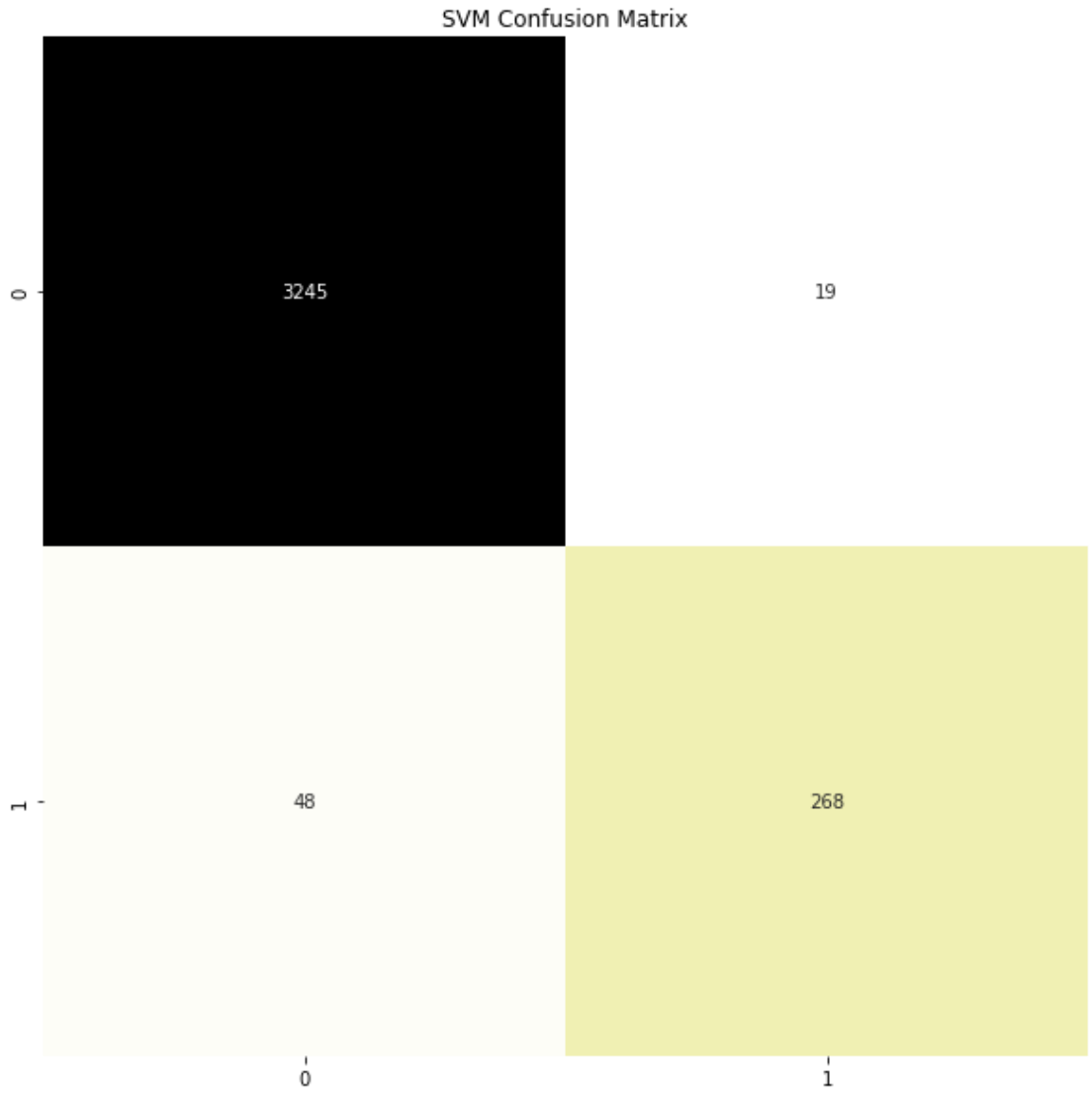
Accuracy: 0.9729050279329609
Recall: 0.7911392405063291
F1 measure: 0.8375209380234506
K-fold Cross Validation Score: 97.27% (0.36%)



Figure[8.]: K-Nearest Neighbours Confusion Matrix

5. Support Vector Machine:

Accuracy: 0.9812849162011174
Recall: 0.8481012658227848
F1 measure: 0.8888888888888888
K-fold Cross Validation Score: 97.93% (0.23%)
`



Figure[9.]: Support Vector Machine Confusion Matrix

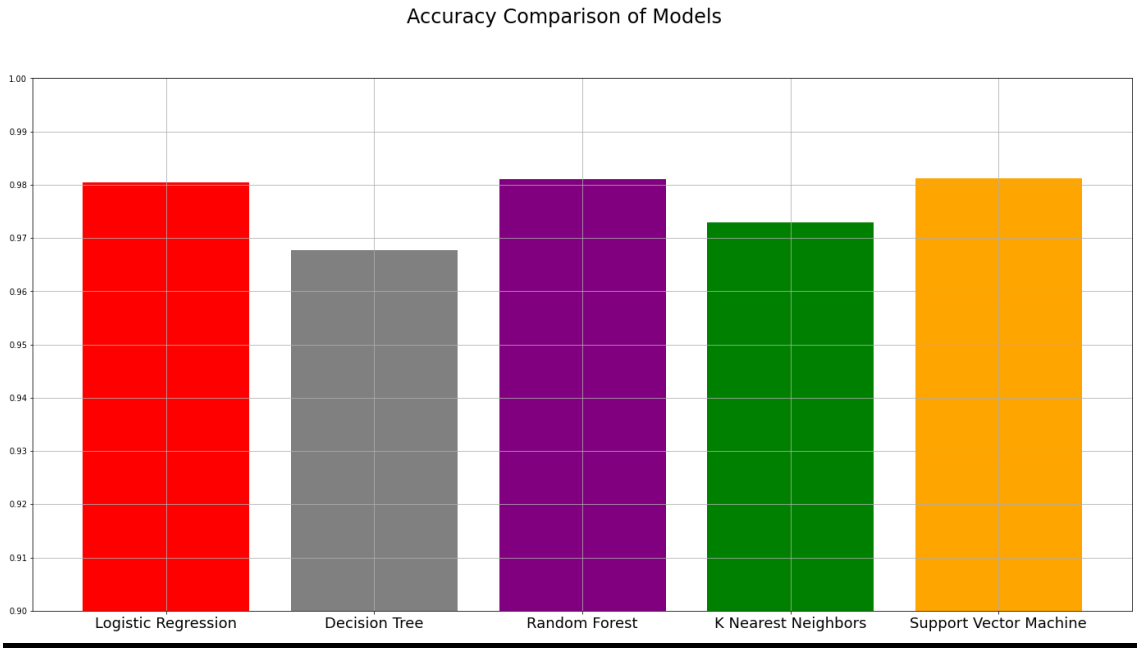


Figure [10.]: Accuracy Comparison of the ML models

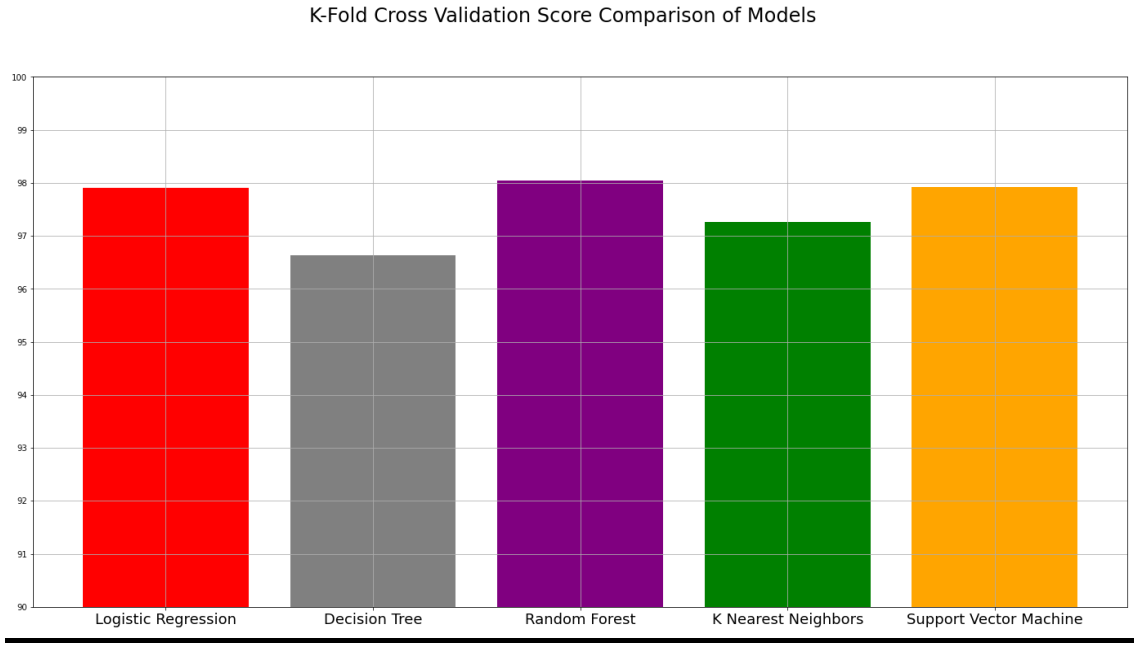


Figure [11.]: K-Fold Cross Validation Score Comparison of Models

CONCLUSION & FUTURE SCOPE

The use of machine learning models for predicting pulsars has shown significant promise in this study. We used five different algorithms: logistic regression, decision tree classifier, random forest classifier, K-Nearest-Neighbours, and Support Vector Machine, to predict pulsars and evaluate their performance.

The results indicate that all the models performed well, with Support Vector Machine and Random Forest Classifier showing the highest accuracy in predicting pulsars, closely followed by Logistic Regression. These models could accurately identify pulsars from the vast amount of data with a high degree of precision.

Machine learning models offer an efficient way to predict pulsars, making it easier to identify these celestial objects, which are crucial to understanding the universe's composition and evolution. The results obtained in this study could help develop more efficient models and algorithms to detect pulsars in the future.

In conclusion, machine learning models offer a significant opportunity for predicting pulsars with high accuracy, and their potential is continually expanding. By using various machine learning algorithms like those we used here, we can create more efficient models and algorithms to detect pulsars more accurately, which can contribute significantly to our understanding of the universe's fundamental nature.

While machine learning techniques show great potential for predicting pulsars, it is essential to acknowledge that they are not a panacea. The success of these techniques depends on the quality and quantity of the data used for training the models, and the choice of algorithm is also critical. Therefore, further research is needed to improve the models' accuracy and the predictions' reliability.

Here are some potential future applications of machine learning in pulsar detection:

1. Improved data processing: Machine learning algorithms can pre-process pulsar data, removing noise and other interference to improve detection accuracy. This can lead to the discovery of new pulsars and improve the accuracy of current detection techniques.

2. Real-time detection: Machine learning algorithms can be trained to detect pulsars in real time. This can help detect short-lived pulsars, which are difficult to detect using traditional methods. The ability to detect pulsars in real time can also help in studying the behaviour of pulsars in more detail.

3. Multimodal detection: Machine learning algorithms can be trained to detect pulsars using multiple modes of data, such as radio, X-ray, and gamma-ray data. This can lead to a more comprehensive understanding of the properties of pulsars and the environments in which they exist.

4. Automated classification: Machine learning algorithms can be trained to classify pulsars into different types based on their properties. This can help study the evolution of pulsars and the relationship between their properties and their environments.

5. Searching for pulsar planets: Machine learning algorithms can be used to search for pulsar planets, which are difficult to detect using traditional methods. This can lead to the discovery of new exoplanets and improve our understanding of the universe.

In short, the future scope of using machine learning in pulsar detection is vast and promising. With the advancements in machine learning algorithms and the availability of more data, we can expect significant advances in the accuracy and reliability of pulsar detection techniques. This research has the potential to contribute to our understanding of the universe and the fundamental nature of matter and energy.

REFERENCES

[1] Vincent Joseph Morello Thesis on discovering Pulsar Stars with machine learning

[2] Hindawl's Predicting pulsars from Imbalanced Dataset with hybrid resampling approach

[3] Kuo Liu introduction to pulsar, pulsar timing and measuring of pulse time-of-arrivals

[4] NSU Florida pulsar search using supervised machine learning

[5] Ieeexplore.ieee.org for Detection of pulsars by classical machine learning algorithms

[6] Machine Learning for everyone

[7] DETECTION OF PULSAR STARS USING MACHINE LEARNING ALGORITHMS by IRJET

[8] Stack Exchange

[9] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

PROJECT TIMELINE

PREDICTING PULSAR STARS USING CLASSICAL MACHINE LEARNING

Company Name: Spartificial

Project Start: Fri, 1-6-2023

Display Week: 1

TASK	PROGRESS	START	END	DAYS
Week 1				
Understanding pulsar star concepts	100%	1-6-23	1-12-23	
Data set collection & description	100%	1-6-23	1-12-23	
Understanding attributes of data	100%	1-6-23	1-12-23	
Finding parameters necessary for prediction	100%	1-6-23	1-12-23	
Refresh ML algorithms	100%	1-6-23	1-12-23	
Reading other research papers on similar problem statements	100%	1-6-23	1-12-23	
Week 2				
Pre-processing Data & handling missing values	100%	1-13-23	1-19-23	
Checking imbalance & solving it through SMOTE	100%	1-13-23	1-19-23	
Learning techniques to handle the imbalanced dataset	100%	1-13-23	1-19-23	
Week 3				
Visualising distributions	100%	1-20-23	1-26-23	
Finding the correlation between variables	100%	1-20-23	1-26-23	
Week 4				
Split the data for training and testing	100%	1-27-23	2-2-23	
Build an ML model	100%	1-27-23	2-2-23	
Measuring various performance metrics	100%	1-27-23	2-2-23	
Generating Confusion Matrix	100%	1-27-23	2-2-23	
Applying K-fold cross-validation	100%	1-27-23	2-2-23	
Week 5				
Review of achieved results	100%	2-3-23	2-9-23	
Final improvisations & implementation	100%	2-3-23	2-9-23	
Week 6				
Going through a few previously written research papers for reference.	100%	2-10-23	2-16-23	
Completing the abstract, introduction and motivation for the project	100%	2-10-23	2-16-23	
Review by the instructor	100%	2-10-23	2-16-23	
Week 7				
Finishing up the literature survey & methodology	100%	2-17-23	2-23-23	
Taking into consideration necessary outputs & graphs for visualisation	100%	2-17-23	2-23-23	
Compiling the final report & inferring conclusions	100%	2-17-23	2-23-23	
Week 8				
Final tweaks while improving the code	100%	2-24-23	3-6-23	
Working on the final format of the research paper	100%	2-24-23	3-6-23	
Getting it reviewed by the instructor	100%	2-24-23	3-6-23	