Roll Number:	**18BCD7143**
Student name:	**L.Satyajit**
Slot (L1/L2/L4):	**L2**
Title of the Project:	**Pulsar Star Prediction**
Objective of the Project (What exactly the project is about?)	**Pulsars are a rare type of Neutron star that produce radio emission detectable here on Earth.Our objective of this project is to predict the pulsar star by the observations from the candidates.**

Dataset Link	Number of rows and Columns	About columns
https://www.kaggle.com/colearninglounge/predicting-pulsar-starintermediate	Rows :9 Columns :12528	Number of Categorical columns: 1 Number of Integer/Float Columns: 8
		Unique Values in each Column: column 1:7192 column 2:12510 column 3:10794 column 4:12528 column 5:7224 column 6:11349 column 7:12526 column 8:11902 column 9:2

Challenges identified in the project	How did you address that challenge?	References
imbalanced dataset	• Random Oversampling • Random Undersampling	https://machinelearningmastery.com/random-oversampling-and-under sampling-for-imbalanced-classification/
missing values	• simple imputer(mean) [Basic] • simple imputer(median) • simple imputer(mode) • Dropping Rows	https://machinelearningmastery.com/handle-missing-data-python/
Noisy Data	• Normalization • Standardization	https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standard ization/

Without Pre-processing- Different Algorithms [Basic preprocessing is done]	Performance(Accuracy/ confusion matrix measures/ROC AUC/ AVG)	Which model worked well on the test data and WHY?
KNN	97.54% / 59F- & 188F+ / 89.59*/97.07%	XG boosting model has a good overall performance compared to any other algorithms as it has good accuracy score , Reasonably minimum false values,Great ROC AUC,and best average accuracy of all.
XG boosting	98.19%/45F- & 136F+/92.46/97.82%	
Support Vector Machine	92.46%/32 F- & 107 F+/92.52%/92.35%	
Random Forest classifier	97.72% / 45 F- & 184 F+/89.88%/97.57%	
Naive bayes classifier	94.36% / 418 F- & 147 F+/89.81%/94.35%	

Which Pre-processing technique you applied?	Why did you apply that pre-processing Technique?	References
dropna()	To Handle missing values	https://machinelearningmastery.com/handle-missing-dat a-python/
SimpleImputer (median)	To Handle missing values	https://machinelearningmastery.com/handle-missing-dat a-python/

SimpleImputer (mode)	To Handle missing values	https://machinelearningmastery.com/handle-missing-data-python/
RandomOverSampler	To Balance Dataset	https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/
Random UnderSampler	To Balance Dataset	https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/
Min Max Scaler	To Handle noisy data	https://www.analyticsvidhya.com/blog/2020/04/featurescaling-machine-learning-normalization-standardization/
Standard Scaler	To Handle noisy data	https://www.analyticsvidhya.com/blog/2020/04/featurescaling-machine-learning-normalization-standardization/

Pre-processing technique name?	Data Mining Algorithm you applied?	Performance(Accuracy/ confusion matrix measures/ROC AUC/ AVG) (Before pre-processing)	(After Pre-processing)	Which model worked well on the test data and WHY?
dropna()	Random Forest Classifier	97.72%/ 45 F- & 184 F+/89.88%/97.57%	98.07%/ 36 F- & 107 F+/ 91.76%/97.98%	KNN works great compared to any other machine learning algorithms. It has a great accuracy score,minimum False positives, excellent ROC AUC values and maximum average accuracy among other models with best
SimpleImputer (median)	Random Forest Classifier	97.72%/ 45 F- & 184 F+/89.88%/97.57%	97.72%/ 45 F- & 184 F+/ 89.88%/97.59%	
SimpleImputer (mode)	Random Forest Classifier	97.72%/ 45 F- & 184 F+/89.88%/97.57%	97.71%/46F- & 184 F+/89.88%/97.59%	
RandomOverSampler	Random Forest Classifier	97.72%/ 45 F- & 184 F+/89.88%/97.57%	93.64%/227F-& 930F+/93.66%/93.5%	
Random UnderSampler	Random Forest Classifier	97.72%/ 45 F- & 184 F+/89.88%/97.57%	94.03%/22F- & 88F+/ 94.09%/93.55%	

Min Max Scaler	Random Forest Classifier	97.72%/ 45 F- & 184 F+/89.88%/97.57%	97.72%/45F- & 184F+/89.88%/97.57 %	preprocessing techniques

Standard Scaler	Random Forest Classifier	97.72%/ 45 F- & 184 F+/89.88%/97.57%	97.72%/45F- & 184F+/89.88%/97.57 %	
dropna(), Random UnderSampler, Min Max Scaler	Random Forest Classifier	97.72%/ 45 F- & 184 F+/89.88%/97.57%	93.82%/13F-& 71F+/93.84/93.16	
dropna()	KNN	97.54%/ 59F-& 188F+/89.59*/97.0 %	97.87%/37F-& 121F+/90.7%/97.26%	
SimpleImputer (median)	KNN	97.54%/ 59F-& 188F+/89.59*/97.0 %	97.55%/57F-& 189F+/89.55%/97.07 %	
SimpleImputer (mode)	KNN	97.54%/ 59F-& 188F+/89.59*/97.0 %	97.57%/57F-&187F+/ 89.65%/97.09%	
RandomOverSampler	KNN	97.54%/ 59F-& 188F+/89.59*/97.0 %	97.86%/383F-& 7F+/97.85%/96.64%	
Random UnderSampler	KNN	97.54%/ 59F-& 188F+/89.59*/97.0 %	92.90%/46F-& 85F+/92.93%/91.05%	
Min Max Scaler	KNN	97.54%/ 59F-& 188F+/89.59*/97.0 %	97.82%/53F-&165F+/ 90.86%/97.53%	
Standard Scaler	KNN	97.54%/ 59F-& 188F+/89.59*/97.0 %	97.90%/50F-&160F+/ 91.14%/97.58%	
dropna(), Random OverSampler, Standard Scaler	KNN	97.54%/ 59F-& 188F+/89.59*/97.0 %	97.82%/288F-&6F+/ 97.82%/96.56%	
dropna()	SVM	92.46%/32 F- & 107 F+/92.52%/92.35%	97.37%/29F-&166F+/ 87.42%/97.34%	
SimpleImputer (median)	SVM	92.46%/32 F- & 107 F+/92.52%/92.35%	97.03%/44F-&254F+/ 86.13%/97.01%	

SimpleImputer (mode)	SVM	92.46%/32 F- & 107 F+/92.52%/92.35%	96.94%/44F-&263F+/ 85.65%/96.91%	
RandomOverSampler	SVM	92.46%/32 F- & 107 F+/92.52%/92.35%	92.19%/293F-& 1128F+/92.21%/92.16 %	
Random UnderSampler	SVM	92.46%/32 F- & 107 F+/92.52%/92.35%	92.03%/34F-&113F+/ 92.09%/91.76%	

Min Max Scaler	SVM	92.46%/32 F- & 107 F+/92.52%/92.35%	97.64%/40F-&197F+/ 89.21%/97.56%	
Standard Scaler	SVM	92.46%/32 F- & 107 F+/92.52%/92.35%	97.87%/40F-&173F+/ 90.50%/97.76%	
dropna() Random UnderSampler Standard Scaler	SVM	92.46%/32 F- & 107 F+/92.52%/92.35%	91.76%/30F-&82F+/ 91.78%/91.25%	
dropna()	Naive bayes classifier	94.36% /418 F- & 147F+/89.81%/94.3 5%	94.58%/303F-&99F+/ 90.38%/94.59%	
SimpleImputer (median)	Naive bayes classifier	94.36% / 418 F- & 147F+/89.81%/94.3 5%	94.33%/425F-&143F+ /89.99%/94.34%	
SimpleImputer (mode)	Naive bayes classifier	94.36% /418 F- & 147F+/89.81%/94.3 5%	94.34%/398F-&169F+ /88.74%/94.32%	
RandomOverSampler	Naive bayes classifier	94.36% / 418 F- & 147F+/89.81%/94.3 5%	89.97%/498F-& 1328F+/89.99%/89.96 %	
Random UnderSampler	Naive bayes classifier	94.36% /418 F- & 147F+/89.81%/94.3 5%	89.48%/57F-&137F+/ 89.54%/89.42%	
Min Max Scaler	Naive bayes classifier	94.36% / 418 F- & 147F+/89.81%/94.3 5%	94.36%/418F-&147F+ /89.81%/94.35%	
Standard Scaler	Naive bayes classifier	94.36% /418 F- & 147F+/89.81%/94.3 5%	94.36%/418F-&147F+ /89.81%/94.35%	

dropna() RandomOverSampler Standard Scaler	Naive bayes classifier	94.36% / 418 F- & 147F+/89.81%/94.35%	90.43%/366F-&923F+/90.44%/90.38%	
dropna()	XG boosting	98.19%/45F- & 136F+/92.46/97.82%	98.41%/37F-&81F+/93.69%98.07%	
SimpleImputer (median)	XG boosting	98.19%/45F- & 136F+/92.46/97.82%	98.16%/47F-&137F+/92.39%/97.84%	
SimpleImputer (mode)	XG boosting	98.19%/45F- & 136F+/92.46/97.82%	98.18%/43F-&139F+/92.31%/97.77%	
RandomOverSampler	XG boosting	98.19%/45F- & 136F+/92.46/97.82%	94.96%/227F-&691F+/94.97%/94.77%	

Random UnderSampler	XG boosting	98.19%/45F- & 136F+/92.46/97.82%	95.34%/22F-&64F+/95.37%/93.49%	
Min Max Scaler	XG boosting	98.19%/45F- & 136F+/92.46/97.82%	98.19%/45F-&136F+/92.46%/97.82%	
Standard Scaler	XG boosting	98.19%/45F- & 136F+/92.46/97.82%	98.19%/45F-&136F+/92.46%/97.82%	
dropna() Random UnderSampler Standard Scaler	XG boosting	98.19%/45F- & 136F+/92.46/97.82%	96.25%/9F-&42F+/96.26%/93.82%	

Summary:

Number of Pre-processing Techniques applied with their names: 8

- dropna()
- simpleimputer (median)
- simpleimputer (mean)
- simpleimputer(mode) ● Random Under Sampler
- Random Over Sampler
- Min Max Scaler
- Standard Scaler

Number of Data Mining Algorithms applied with their names: 5

- KNN
- XG boosting
- Random Forest Classifier
- Support Vector Machine
- Naive bayes classifier

Which algorithm showed highest performance after all pre-processing techniques and WHY?:

KNN works great compared to any other machine learning algorithms.It has a great accuracy score, minimum False positives, excellent ROC AUC values and maximum average accuracy among other models with best pre-processing techniques .

Conclusion-Write in your own words:

As we observe, XG boost worked well before pre-processing but had a lot of false positives. In our experiment false positives has higher priority than false negatives as it would spoil the integrity of our experiment. After all pre-processing SVM model brought out very good figures with minimum false positives, which is indeed a solution with improved precision.