ASSIGNMENT
ON
Machine Learning

Submitted By:

Jyothis Benny

20PMC135

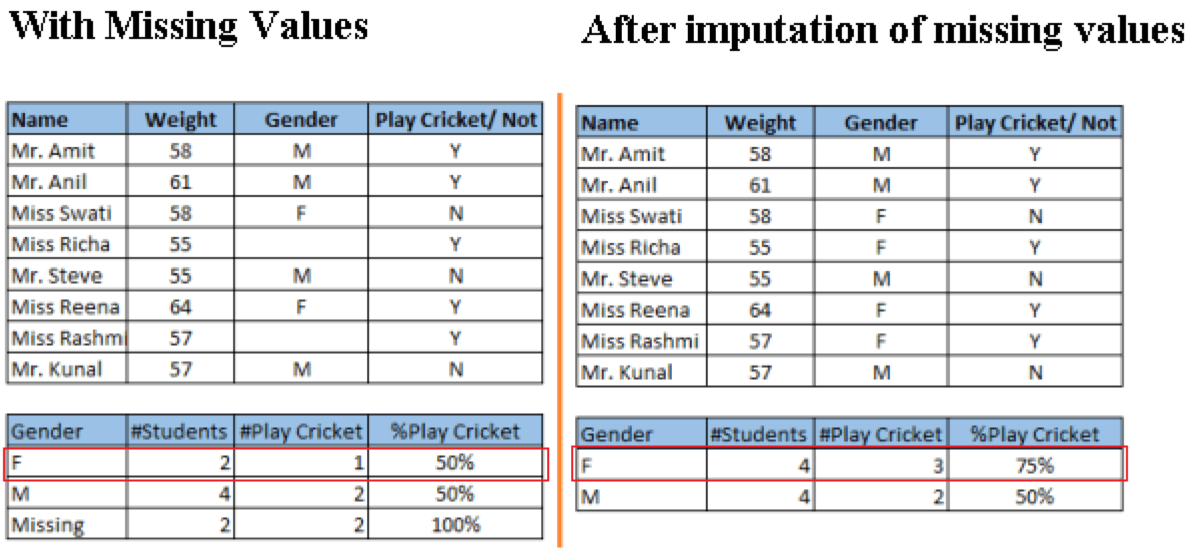
➔ Identify the techniques to improve the classification accuracy.

1. Add more data

Having more data is always a good idea. It allows the "data to tell for itself," instead of relying on assumptions and weak correlations. Presence of more data results in better and accurate models.

2. Treat missing and Outlier values

The unwanted presence of missing and outlier values in the training data often reduces the accuracy of a model or leads to a biased model. It leads to inaccurate predictions. This is because we don't analyse the behavior and relationship with other variables correctly. So, it is important to treat missing and outlier values well.



With Missing Values

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55		Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57		Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	2	1	50%
M	4	2	50%
Missing	2	2	100%

After imputation of missing values

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55	F	Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57	F	Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	4	3	75%
M	4	2	50%

Above, we saw the adverse effect of missing values on the accuracy of a model. Gladly, we have various methods to deal with missing and outlier values:

1. **Missing:** In case of continuous variables, you can impute the missing values with mean, median, mode. For categorical variables, you can treat variables as a separate class. You can also build a model to predict the missing values. KNN imputation offers a great option to deal with missing values. To know more about these methods refer article "Methods to deal and treat missing values".

2. **Outlier:** You can delete the observations, perform transformation, binning, Imputation (Same as missing values) or you can also treat outlier values separately. You can refer to the article "How to detect Outliers in your dataset and treat them?" to know more about these methods.

3. Feature Engineering

This step helps to extract more information from existing data. New information is extracted in terms of new features. These features may have a higher ability to explain the variance in the training data. Thus, giving improved model accuracy.

4. Feature Selection

Feature Selection is a process of finding out the best subset of attributes which better explains the relationship of independent variables with target variable.

You can select the useful features based on various metrics like:

- **Domain Knowledge:** Based on domain experience, we select feature(s) which may have higher impact on target variable.
- **Visualization:** As the name suggests, it helps to visualize the relationship between variables, which makes your variable selection process easier.

5. Multiple algorithms

Hitting at the right machine learning algorithm is the ideal approach to achieve higher accuracy. But, it is easier said than done.

This intuition comes with experience and incessant practice. Some algorithms are better suited to a particular type of data sets than others. Hence, we should apply all relevant models and check the performance.

6. Algorithm Tuning

We know that machine learning algorithms are driven by parameters. These parameters majorly influence the outcome of learning process.

The objective of parameter tuning is to find the optimum value for each parameter to improve the accuracy of the model. To tune these parameters, you must have a good understanding of these meaning and their individual impact on model. You can repeat this process with a number of well performing models.

This is the most common approach found majorly in winning solutions of Data science competitions. This technique simply combines the result of multiple weak models and produce better results. This can be achieved through many ways:

- **Bagging** (Bootstrap Aggregating)
- **Boosting**