Before, working on any ML model it is very important to gain insight about the model and dataset which you are using to automate/solve any problem, which can be easily done but not done my many developers is **hypothesis space,** that is, generating, exploring and impact of features/attributes/hypothesis on which you are working.

But how does it help? It helps to remove/lower bias and variance made by model, which provided to it by our data or any of its feature. This is the most crucial step to improve performance/accuracy of any ML model. Let's discuss some possible aspects

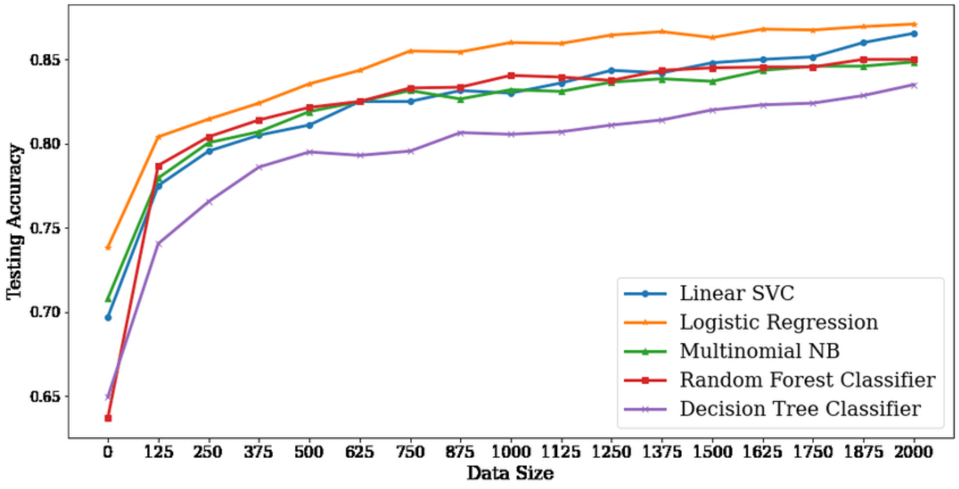
Methods to Improve Accuracy of any ML Model:

1. Add more data

Having more of data points in your dataset is good, it provides:

- Good assumption ability to our model
- Improve weak correlation between features

It's very difficult to increase the size of a dataset when you are working on any project but when your existing model accuracy is low then you must struggle to collect more data, to make it more familiar with your problem-statement.



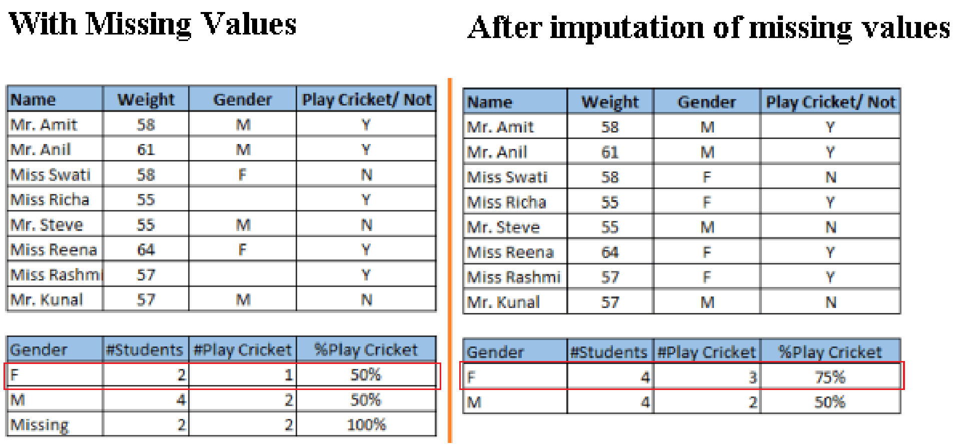
This above depicts different models' accuracies with respect to size of dataset.

2. Fill Missing/NaN values

In dataset, there can exist many data cells which are empty, that is, do not contain any value which eventually lead to wrong prediction of our dataset or increase biasing. All this happens just because our model is unable to analyze the relationship between variables.

So, it is necessary to fill them it can be done in many ways:

- If data is continuous, fill it with its corresponding columns mean, median or mode value.
- If data is discrete fill with most occurring value in its corresponding column.
- If they contribute to very less percentage to our dataset, then remove them.
- You treat them as separate class/remove it from training dataset and then after training our model, with the help of parameters you find their value.
- Fill them with any value and add new column containing binary values, 0 when data is not filled by yourself else fill 1. They add extra weight while optimization.



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%

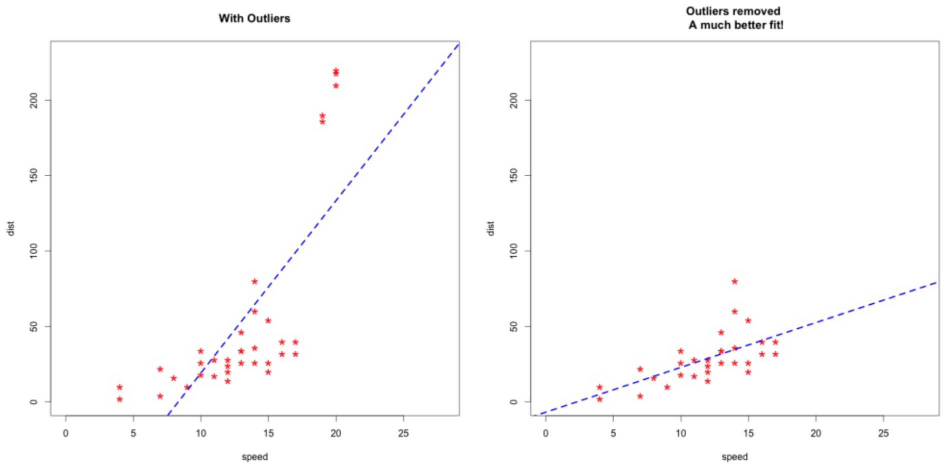
After imputation of missing values

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%

3. Treat Outliers

Remove all unwanted/wrong data from your training dataset because their work is only just to misguide your model to reach optimal value, make wrong predictions/decisions while testing. You can identify them by exploring and analyzing your data. If you want you can use KNN/SVM like instance-based classification models to identify outliers (how, data point which is very less crowded/far away hyperplane).

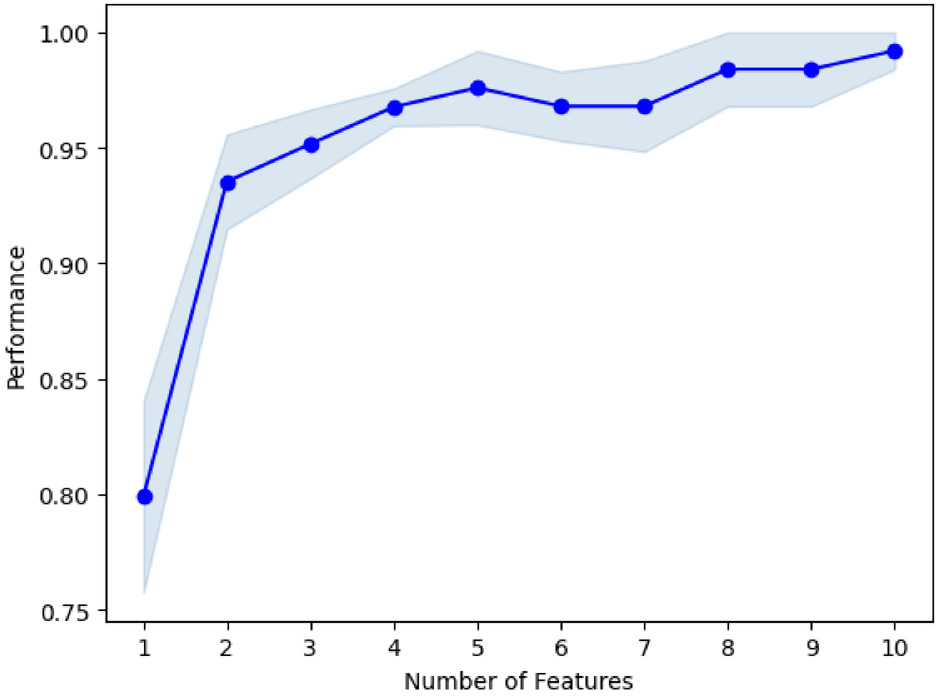


The above graph depicts, impact of outliers on regression best-fit line.

4. Feature Selection

Features which are relevant to our model and problem statement should only be present in our dataset because they majorly contribute to model performance. Irrelevant features just add extra weight/bias/variance to make wrong prediction and to never converge over global minima. To find relevant features:

- Find their correlations with respect to dependent variable, less correlated feature should be discarded.
- Remove all features from dataset and then add one by one in dataset such that at each step you check your model overall accuracy increasing/decreasing, according to that accept/reject them.
- Select according to domain knowledge.
- Visualize the relationship between features and dependent variables in graphical form.
- Some popular techniques:
    - Forward selection
    - Backward selection
    - Bi-directional search
    - Backward propagation
    - Genetic Algorithms
    - PCA

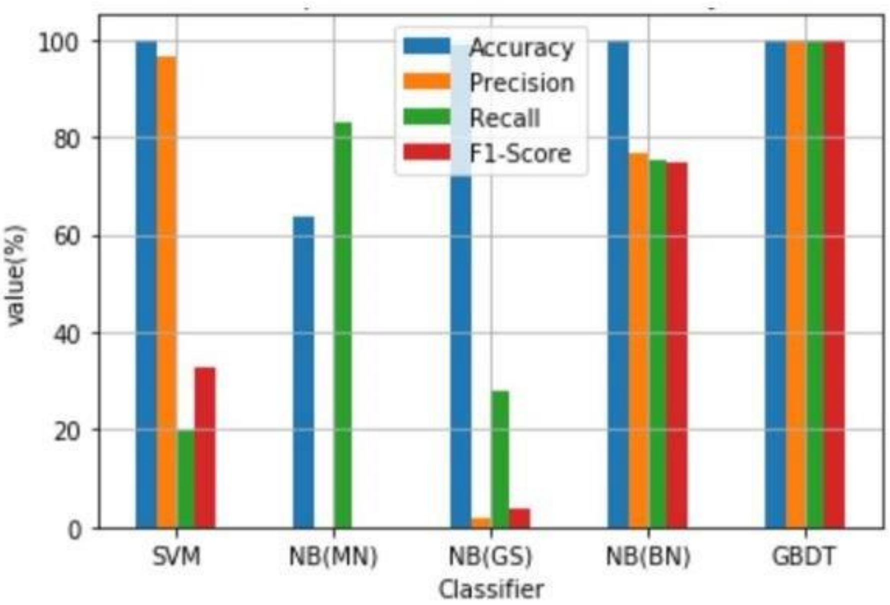


5. ML Algorithm/Model Selection

It is not necessary that to solve a given problem all ML Algorithms would perform the same because your dataset may not support every algorithm/model requirement.

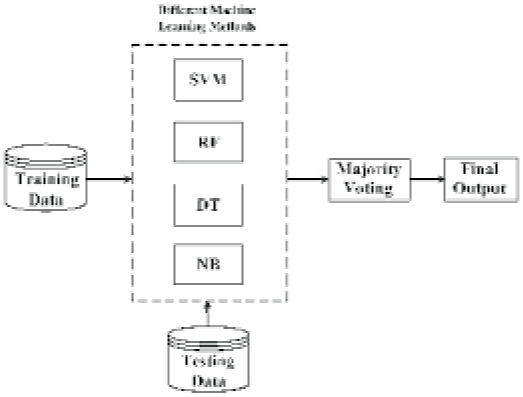
For example,

- Naïve Bayes requires independent feature dataset so, it should be chosen where features are highly correlated.
- Logistic/Linear Regression requires features to be linearly dependent on each other so, this should not be chosen when data related in any polynomial/exponential fashion.
- KNN assumes that highly similar data points are very close to each other.
- Simple SVM requires data to be linearly separable by just a hyperplane or linear line with appropriate slope and intercept value. So, it should not be used where data points are separable by complex boundaries.



6. Ensemble Learning

Combining two or more different machine learning model can help us to increase the accuracy because limitation of anyone's model is resolved by another model. To achieve this idea there are many populare techniques available such bagging (aggreating), boosting, stacking and random forest etc. But this turn our simple and basic model into complex method and choosing a approriate base model should be done carefully.

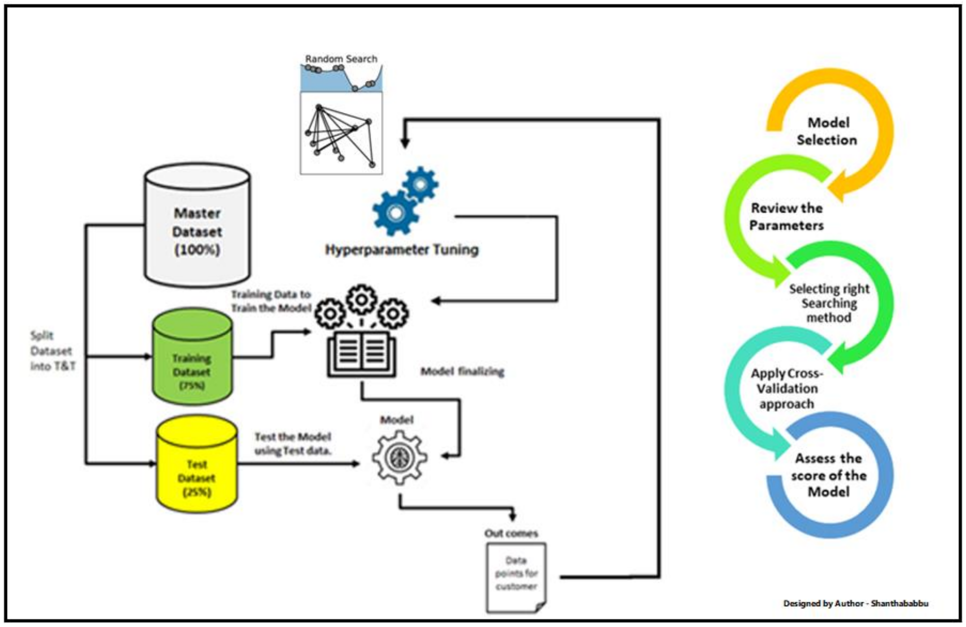


The above diagram depicts an example of ensemble learning technique.

7. Hyperparameter Tuning

Learning of model while training is highly influenced by hyperparameters and their values. So, it should be highly optimal to improve accuracy which can be found out by seeing their meaning and impact on the model. They can be tuned by some AI techniques as well

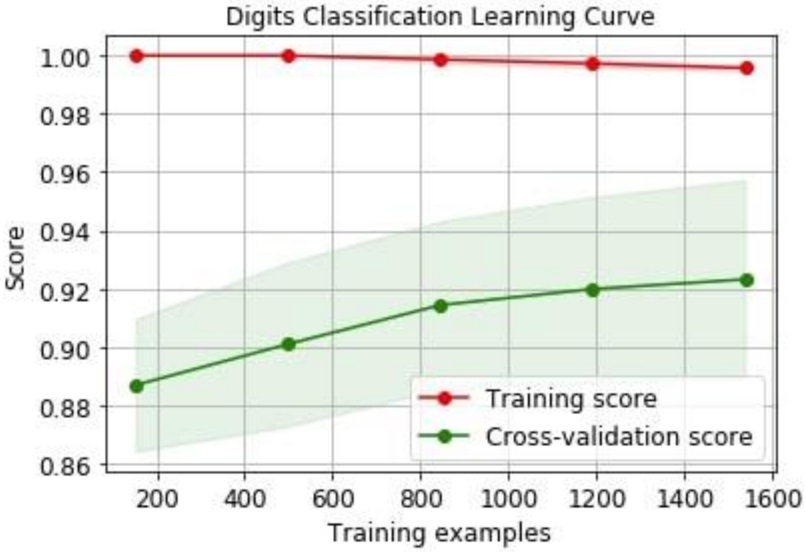
- Simulated annealing
- Hill climbing
- Genetic Algorithm and its variants



The above diagram depicts a flowchart for tuning hyperparameter.

8. Cross Validation

It is the concept of data modeling, which says that try to leave a sample dataset for future testing and train on the rest of dataset before finalizing your model.



For this method we need to split our dataset into 3 different parts called training, validation and testing. Most popular technique in cross validation is K-Fold.

Conclusion:-

All these above listed methods works well when you want to increase the accuracy of the model but along with these things you should have knowledge of multiple

1. Artificial intelligence and deep learning methods through which you can converge your model's cost function to global minima with appropriate learning rate and epochs.

2. Machine learning algorithms to always have an alternative for a given problem or you can experiment with different algorithms.