Preprocessing :

Is an important process in Data Mining and Machine Learning projects as it the process which the Garbage get out from the dataset so it's the cleaning process where inaccurate records are removed or corrected this is Preprocessing in general but,

Specifically Preprocessing for images it's not important it's necessary step as images at the lowest level of abstraction so preprocessing works for improvement important features in image data for further processing.

Image Preprocessing changes according to size of pixels used to predict the new ones

In our project all models use the same Preprocessing steps to prepare dataset for training and then testing. our dataset consist of 2-folders ("Uninfected "= don't have malaria, "Parasitized"=have malaria)

Preprocessing Steps :

1- we took files name in each folder
2- we want to remove file called ("Thumbs.db ") from both folders
3- now we have all images of both classes Uninfected and Parasitized but in different pixel rate so, we need to resize all images to be the same and change the color mode to "L" for all images in both classes
4- here we want to graph images in X(Pixels) and Y(Classes)
5- now images converted into numpy array and ready for training and testing
6- so we split the data to train 70% , test 30%

Modeling:

1- Naïve Bayes :

Is one of the Supervised ML classification algorithm . it works on Bayes Theorem of probability

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{B}|A) \cdot P(A)}{P(B)}$$

to predict the class of the unknown dataset

P(A|B) =>is a post probability of a class.

So, How Naïve Bayes algorithm works ?

It calculate post probability (P(A|B))of all classes and the class of high post probability comes to be prediction in test data

But like anything in our life their is Pros and Cons so Naïve Bayes do

Pros	Cons
It's easy and fast to predict test data	If their is a category in test data that is not observed in training data it failed to get a prediction and assign zero to probability
Works better in independence feature than LR	It's too difficult in real life to find an independence features
	It called bad estimator

That 's a general talk about Naïve Bayes and why use it as a Classification algorithm and after using it

It gives us an accuracy =51 at Random State=100

And when we decrease Random State to be =2 the result accuracy = 50

2- Random Forest :

It's a Supervised ML used for both Classification and Regression .Forest build many of decision trees on randomly selected data sample then getting prediction from each tree and taking means of voting as the prediction for test data

Random Forest algorithm :

1- select random sample from dataset
2- create Decision Tree for each sample and get prediction result from each tree
3- make a vote for each prediction
4- select the most voted prediction as a final value for test data

here the advantage and disadvantage of Random Forest

Pros	Cons
Highly accurate and robust model	Slow in generating prediction as no. of trees raise
Doesn't suffer from overfitting problem	Model is difficult to interpret compared to decision tree
Can be used for both Regression and Classification problems	
Can handle missing values in dataset	

Random Forest technically an ensemble method based on divide and conquer technique as it a collection of decision trees

For all Random Forest privilege we choose it to be our second model

So we get accuracy = 81 when the no. of estimator(trees) = 20

And even that we increase no. of trees to 150 the accuracy remain the same

Ensemble Learning :

After we apply two different models on our dataset now it time for ensemble learn but, First what's The Ensemble Learn ?

> It 's the process which uses multiple machine learning models are constructed strategically to solve a problem. It attempt to improve the performance of model by enhancing the Accuracy of the model and prevent over fitting

Then we need to know How Ensemble Learn work ?

Concept of ensemble learn is to train multiple models on the same data and combine their prediction into only one output

There is different ways to apply ensemble learning. It can be by Voting ,bagging ,Random forest all these are techniques for implementing ensemble learn

To practice Ensemble learn in our project we perform Bagging Technique on Naïve Bayes as it has low accuracy so we want to increase it and Naïve Bayes is relatively fast so after applying bagging it wouldn't than bad nor slow

After applying bagging the accuracy increased from 50 to 60 which not bad and also it didn't take to loge to get result so it's quite acceptable

Conclusion:

As we use Ensemble learn the Accuracy increased but also it takes a lot of time to train the model and get a result so in applying ensemble learn you had to chose technique that matches with your data and your preprocessing to increase the accuracy as much as possible

As it is important to chose which technique to use in Ensemble learn also your Preprocessing affects the accuracy so bad so you to know which technique to follow in preprocessing according to your dataset

Finally their no perfect technique you need more accuracy then you have to wait for it and the tip is to balance between efficiency and Time