This is the assessment report prepared to demonstrate my skills required for the 3D Computer Vision Robotics Research scientist role.

Short note on task to be done.

Have been given two different datasets. one obtained from an apple tree which includes data on hundreds of apples. Another data obtained from 245 trees. This may or may not contain apples. Given the Cartesian coordinates (X, Y, Z). find the way of finding/count the number apples.

Implementation

My focus for this task is mainly on the datasets under some apples folder, each txt file contains 12 columns. In which the last 6 columns contain non-essential data, as we had to ignore them(as mentioned on the brief), they were dropped using the drop command for the process of cleaning the data, also we don't want the robot to get stuck trying to perform the tasks, which will be difficult/improper for it, when the data is corrupted (I.e., missing column values, partially occluded data), as might cause delay in performing the task quickly, we will drop these type of data as well from the csv file. Another thing would be, when the data is collected from a partially occluded fruit, we won't be able to estimate the volume(size) of the fruit, this could cause a 150mm sized fruit to get down to the range of our apple's size. Finally, the unlabelled data in the csv files were labelled while reading the csv files (naming them aptly).

As hinted on the brief, the apples were classified among the other fruits, by using their cartesian coordinates to find the fruit size, comparing it with the mentioned apple size and upon some research through the web, I got to know that I can use Naïve Bayes classification algorithm to find if the fruit is spherical. So, if these two criteria satisfy, the fruit must be an apple (looking at the information provided from the brief), so saved them in a different csv file, this can be provided to the robot to grab the apples later. I thought of using the average of colour intensity to predict the fruit's colour but later discarded that approach, as this would increase the complexity of the model leading to accuracy of prediction demining or/and the robot picking the wrong fruit instead of apple. Overfitting could even cause delay in prediction; the process needs to be quick in our case.

Another reason for not using the R, G, B dataset provided for classifying apples is due to lack of information provided about the Farm where the fruits are being harvested, what variety of apples are being harvested (there are around of 7500 varieties), as each variety of the apple changes in its nature (size, colour & probably even a bit in shape).

In order, to find the size of the fruits, the formula to find the fruit's volume is the closest thing possible to finding its size. Assuming that all of the fruits would be almost spherical, because there isn't much information provided on the brief about the fruits being harvested in the farm. As, we're sure here that there would be at least one spherical fruit, decided it will be a good idea calculating the formula to find the volume of a sphere on the data provided. Need to find the radius using the cartesian coordinates. Upon some research, found that getting the square of each coordinate, as they were in meters, converted to millimetres (but this didn't work so scrapped the idea of conversion) and adding them together and finally taking a square root of it would give the radius of the sphere. After finding the radius using this equation, used it in a formula to find the sphere's volume multiplying 4/3 with pi and the radius.

After doing my research on the internet about finding an object shape through just the X, Y, Z, R, G, B, data provided for this task, although it could probably be done using Naïve Bayes Classification model could probably increase the complexity leading to overfitting or would need more time to be

spent on the research compared to the expected task complexity, when using colour to classify apples from the other fruits. From this, I concluded that some more data and possibly some more time might be required to find the shape of the fruit. Due to this, I changed my plan for implementing this task, which is why I haven't removed my plan earlier from this report to demonstrate that I adapt to the changes quite quickly, by thinking of other possibilities of implementing the task well.

Now, I have done my research on finding the colour of the fruit to improve identify, because identifying apple, just on its size wouldn't be accurate (there are approximately 100 fruits sized in the range, while checking the condition on a cleaned dataframe). I saw that there are majorly 4 types of colour apples, Red, Yellow, Green and white. My plan now to find the apples accurately would be, if the fruit size is within the threshold (between the 60 to 100 mm) **AND** any of the 4 colours, it must be an apple, although, this wouldn't be much accurate. With the given data, this could be the best possible way to find the apple.

In order to find the fruit colour, I have assumed that the apple harvested must either be one of the major apple colours: Red, Green, Yellow & White (although these apples are rare, didn't want to neglect them) I have used a threshold (as, the apple colour might differ in ripeness or other factors) value for the colour codes. I have defined a separate method for this task. Which gets the column data (RGB) from the csv file, and then find the average colour and compare if it's one of the fruit values with regards to the threshold.

I am just a recent graduate enrolled in a machine learning certification course, in the process recently started working with a project on machine learning, which is to mimic the working of Facebook marketplace recommendation ranking system. I am not that familiar with these tools and syntax(pandas), Currently the script has an error, which needs to be corrected. Performed mathematical operations on a numerical string (Data saved in the csv file is a string) later to find out that isn't possible in python, resolved this issue using type casting, but later encountered another error which seems to persist, I am pretty sure, this is due to storing the dataframe into the csv file, as instead of storing the data as a table they are being stored as a list of strings, indexing and adding column names didn't work. Would need some more time for research, inorder to get the script working. till now haven't started working with deep learning tools, which might have been handy in performing this task a lot better. but I can assure, you that I am good with general programming, you might have seen the robustness and following the good coding practices. I am pretty sure, that I'll get familiar with the tools and syntax before joining the work.

References

https://machinelearningmastery.com/classification-as-conditional-probability-and-the-naive-bayes-algorithm/

https://www.youtube.com/watch?v=3I8oX3OUL6I

https://www.datasciencemadesimple.com/square-root-of-the-column-in-pandas-python-2/

https://www.researchgate.net/figure/The-RGB-minimum-and-maximum-value-colour-codes_tbl2_349549178

https://www.digitalocean.com/community/tutorials/python-convert-string-to-float