MACHINE LEARNING TEST TASKS

ML1) Ways to handle imbalanced datasets:

- **Collecting more data:**

A larger dataset might expose a different and perhaps more balanced perspective on the classes. Collection and addition of more data to the minority class can lead to a balancing of the dataset.

- **Changing performance metric:**

Looking at the following performance measure can give more insight than the tradition classification accuracy of the model:

Confusion Matrix: A breakdown of predictions into a table showing correct predictions (the diagonal) and the types of incorrect predictions made (what classes incorrect predictions were assigned).

Precision: A measure of a classifier's exactness.

Recall: A measure of a classifiers completeness

F1 Score (or F-score): A weighted average of precision and recall.

- **Resampling of the dataset:**

Adding copies of instances from the under-represented class called oversampling

Remove instances from the over-represented class called under-sampling.

I recommend trying both of them on the imbalanced datasets which can maybe lead to an increase in the performance.

- **Using different algorithms:**

Different algorithms should be tried to different scenarios instead of using one algorithm for all the scenarios.

ML2) To solve the problem mentioned in this task we can use the following approach:

- **Reading CSV files in chunk:**

When CSV files are read by specifying chunk size then the original data set is broken into chunks and stored in pandas parser object. Then we iterate the object and concatenate to form the original dataset that takes lower time to process it.

This can help in reading large datasets in less time and avoid system crashes.

- **Data type size should be changed:**

The processing time of large datasets with a huge amount of space is more. To avoid this we can change the datatypes of certain columns like (int 64 -> int32),(float64 ->float32) to reduce the space it stores, and then save it into a CSV for further use.

- **Removing unwanted columns from the dataset:**

Removing unwanted columns can lead to a decrease in the storage space of the dataset. We can find the unwanted columns using the heatmaps which gives us the weightage of each column to the result.

- **Changing the data format:**

If we are having the dataset in CSV format, we can convert it to different data formats such as GRIB, NetCDF, or HDF to speed up data loading and use less memory.

- **Calculating the math efficiently:**

Using NumPy and its functions can help in optimizing and reducing the run time of the datasets. Numpy should be used instead of python math and NumPy arrays should be used instead of python arrays.

Also using map and reduce functions.

- **Vectorization of functions:**

Instead of using nested for loops vectorization helps in optimization.

If we compare for loops approach and vectorization approach, we can see vectorization is more efficient and less time-consuming.

- **Using n_jobs parameter in sklearn:**

N_jobs parameter in sklearn tells sklearn how many jobs to run in parallel both to training and transformations. This will help in saving time during the processing of the dataset.

ML4) The variables that need to be chosen should have a direct impact on the preferred career of the student.

For example, The address, phone number, email, etc are not important for career guidance decisions and hence can be neglected from the dataset containing the details about the student.

According to me, the following important variables need to be included for career guidance of a student:

1.Highest qualification:

This variable lets us know at which level the student is eligible to work according to which the person can opt for their career.

For example, If the highest qualification of the student is B-Tech and that of another student is M-tech, this would indicate that the second student is eligible to work at a higher level than the first.

2.Interests:

This variable should match with the profile of the preferred career choices of the student to enjoy and excel in their work simultaneously.

3.Marks:

This will tell about the technical skillsets of the student. This will help in specifying the type of job the person should do.

Ex: If a CS graduate has more skills and interests in website designing than electronics and hardware then he/she should be guided accordingly.

4.Personality and soft-skills:

This variable tells us whether the student is suitable for their preferred career choices. This variable should also match with the profile of their career choices.

Soft skills include analytical, communication, collaborative skills.

Other skills include: supervising, managing, influencing/persuading, counseling, consulting, problem-solving, decision making, organizing, planning, creating, computing.

Ex: If a student is introvert and lacks basic skills in communication then there is no point in having a career choice where they need to address people a lot.

Concluding this topic, I want to say,

Skills give you a job, Skills with interest makes you a career, Skills with interest meaning and purpose-vocation

YUDHAJEET BHATTACHARYA

(yudhajeetm@gmail.com)