Machine Learning Assignment

Instructions: Read carefully !!!!!

There are two types of questions. Theory and coding. Please note that while answering the theory part, you cannot copy paste anything from internet / AI tool like chatgpt. Your script will be checked by using Turnitin and if any copy/paste is found in the theory part then you will not get any marks. So please don't copy paste anything but you can read article related to your topic and then you can use those knowledges to write the answer of the theory part. In the coding section you're free to use code from Github/Kaggle etc. But note that you have to write proper comment (exactly to the point) for each and every line of your code. Note that if you are using exactly the same code (it must be 100% same not even 99.99%) again then you don't need to write the comment again but if you want to write you can do it. As you're entering in the field of Machine Learning and research, you should be comfortable with your keyboard. Because in future you will write a lot of papers / articles / codes. Pen and paper is good for practicing but after all you will not be able to use pen and papers in any official submission. So please write your answer in MS word / latex for the theory part. MS word is good enough for you now. If you are not familiar with latex then ignore it for now. Both latex and MS word answers will be treated equally. Also, grammar related mistakes will be ignored while marking. But try your best to minimize grammar & spelling mistake. If you want to add any images/graphs/figure, then for now, draw it in a page and then take a picture and add it to your MS document. Make sure your image is properly aligned with your answer. If someone want to make images/graphs/figure in excel / power point / in any software, then it is definitely okay and it is a good practice for your future. Also finish your writing within the given word limit. It is sufficient enough for explaining.

----------Theory part----------

Q1. Assume that while training a model, we have 60% accuracy on training dataset (without cross validation) and 90% accuracy on test dataset. Do you want to use your model in real life application? Explain your answer. *Marks - 5 (Word limit 200)*

Q2. Imagine you are working on a very large tabular dataset. For example, 1000000 samples. Now you have found that one of your feature F1 has 1000 missing values and another feature F2 has 80000 missing values. How you will handle it? There can be multiple ways. Any logical answer will be accepted. *Marks - 5 (Word limit 200)*

Q3. Ther is a famous series called “Lord of the Rings”. In that series, there are two warrior nations called the elves and the dwarfs. The elves are usually tall and light weight and the dwarfs are short and have more weights than the elves. We have a dataset that have two features height and weight. Now we would like to use a classification process to determine that if a person is elf or dwarf based on height and weight. Consider this entire dataset as training data.

Height	Weight	Class / target / outcome
7	40	elf
5	50	dwarf
7.1	42	elf
7.2	42	elf
5.1	52	dwarf
5.2	55	dwarf
4.8	50	dwarf

(a) Make a Decision Tree use both features, calculate the Gini value and find the leaf nodes. No theory explanation is required. Just show the calculations. And draw the tree step by step (multiple figures) *(10 marks) (Word limit -)*

(b) Sow how you will use KNN algorithm on this training dataset for training. Again no need to explain theory. Just show the steps. *(5 marks) (Word limit -)*

(c) In part a and b we have trained a Decision Tree and KNN for a classification problem. Now We have a new data (test data) which features are Height = 7 and weight = 50. So, the task is to find the class / nation of this new datapoint by using both of your model. You must show the steps. *(5 marks) (Word limit -)*

(d) Now here is the critical part. The new data (test data) given previously (Height = 7 and weight = 50), Consider that the actual class of that data is “elf”. Now see if both of your classifier are giving correct prediction or not. If anyone of the classifier is giving wrong prediction then try to find a reason behind the wrong prediction and how can we solve it? Note that we cannot increase the amount of training data and also we cannot change our models. *(10 marks) (Word limit- 500)*

----------Coding Part----------

(Q1) In the tutorial part you have seen the housing price prediction problem. Your job is to improving the performance. Means you have to minimize RMSE, MSE, MAE and maximize R2 value without overfitting. You can use any model. Initially for your benchmark you can consider that your target is to make the MAE value 1000. This is just an assumption that if my model predict 1000 taka wrong then I am happy to consider it. But remember that your task is to minimize the RMSE, MSE, MAE and maximize R2 value as much as you can. Consider this as your first step to paper publishing. When you will try to publish a paper, in most of the cases, you have to beat the current state of the art somehow. *Marks- 30*

(Q2) Previously you have seen how to classify diabetes patients. Now you have another diabetes dataset where the target is either 0 or 1. 0 means no diabetes / No and 1 means the patient has diabetes / Yes. Now your task is to start from the data preprocessing to make prediction step by step. Note that in this problem, your main target is not to make a good accuracy. The main focus will be how you will handle this dataset from the pre-processing to the training the models. Please write down your steps in a markdown cell so that I can understand what you have done. Please be clear that without proper explanation, only writing code will not help you to get a good mark. For this question always mention two things “what are you doing?” and “why are you doing this?”. Accuracy is the second priority here. *Marks- 30*