



IITK Consulting Group

Secretary recruitment tasks - Data Science domain

Instructions

1. The tasks are separated into three categories since each learner has a different amount of interest and familiarity with various ML/DL subdomains.
 2. You are allowed to submit **any one** (or more) problem statements in any bucket you like
 3. You must submit the predictions file on Kaggle for your predictions to be judged because all PS are either active or expired Kaggle contests.
 4. We've assigned specific points to each PS. Therefore your score for a given PS will be calculated as **(Points*Accuracy)**, where Accuracy is the score obtained by Kaggle when you submit your predictions.
Example: If a PS has 40 points associated with it and you score 86% accuracy on Kaggle, then your score will be $40 * 0.86 = 34.4$
 5. A weighted average of all the individual task results will be used to determine your final recruiting task score.
Example: If you get 38 (out of 40) and 42 (out of 50) on two different tasks, your final score will be 88.88. If you just undertake one task, the results of that task's evaluation will be taken into account
-

Bucket 1 - Data analysis + Machine Learning

Problem statement 1

40 points

Link: <https://www.kaggle.com/competitions/spaceship-titanic/overview>

Evaluation criteria: 15 points for Data Analysis + $25 \times \text{Kaggle_Accuracy}$

Comments: We expect informative data analysis using charts and graphs. For this purpose, you are free to explore libraries, including Matplotlib and Seaborn.

Problem statement 2

45 points

Link: <https://www.kaggle.com/competitions/walmart-recruiting-trip-type-classification/overview>

Evaluation criteria: 20 points for Data Analysis + $5 \times (5 - \text{Kaggle_score})$

Comments: We expect *very strong* data analysis using charts and graphs. For this purpose, you are free to explore libraries, including Matplotlib and Seaborn. **In** this task, Kaggle calculates Logarithmic loss, so the lower the score, the better it is. Anyone with a Kaggle score > 5 will be considered disqualified for this task (highly inefficient model).

Bucket 2 - Deep Learning and its applications

Problem statement 3

60 points

Link: <https://www.kaggle.com/competitions/plant-pathology-2021-fgvc8/overview>

Evaluation criteria: 20 points for Data Analysis + $40 \times \text{Kaggle_score}$

Comments: We expect efficient pipeline modelling along with basic data analysis charts. You are free to use libraries and frameworks like Tensorflow, PyTorch, and Sklearn.

Problem statement 4

50 points

Link: <https://www.kaggle.com/competitions/dog-breed-identification/overview>

Evaluation criteria: 20 points for Data Analysis + $15 \times (2 - \text{Kaggle_score})$

Comments: We expect efficient pipeline modelling along with basic data analysis charts. You are free to use libraries and frameworks like Tensorflow, PyTorch, and Sklearn.

In this task, Kaggle calculates Logarithmic loss, so the lower the score, the better it is. Anyone with a Kaggle score > 2 will be considered disqualified for this task (highly inefficient model).

Bucket 3 - Natural Language Processing

Problem statement 5

75 points

Link: <https://www.kaggle.com/competitions/feedback-prize-effectiveness/overview>


Evaluation criteria: 25 points for Data Analysis + $10 \times (4 - \text{Kaggle_score})$

Comments: We expect *very efficient* pipeline modelling and fundamental data analysis. You are free to use libraries and frameworks like Tensorflow, PyTorch, SpaCy, and Hugging Face. In this task, Kaggle calculates Logarithmic loss, so the lower the score, the better it is. Anyone with a Kaggle score > 4 will be considered disqualified for this task (highly inefficient model).

Submission

- Form: <https://forms.gle/VY8YVHts8ohPqBPd6>
- **Deadline - 21st July 2022, 11:59 PM**
- Create just one notebook containing the data analysis and modelling for a single task. Submit all the notebooks and relevant screenshots in a single form.
- Notebook naming convention - **TaskNumber_FirstName_RollNumber.ipynb**
Example: For task 3, the submission will be **3_Shreyansh_200953.ipynb**

Relevant resources

1. Page 4 of  ML roadmap **(MUST DO)**
2. An introductory article on [Data analysis](#)
3. Introduction to [ANNs](#)
4. Best [course on ANNs](#) to binge
5. Introduction to [CNNs](#)
6. TensorFlow [crash course](#) (High enthu required XD)
7. [TensorFlow documentation](#) to understand the implementation of CNNs
8. All about [Natural Language Processing](#)

Expectations and personal advice

Everything is accessible via the open internet and only takes a click. Keep in mind that Google is your most excellent pal! Use the materials offered wisely, and don't be afraid to look for any intermediate hyperlinks that appear. Finally, you must be able to defend your written code and explain your strategy to us. Scores don't matter; they only serve to measure information. People that are willing to devote all of their time to learning something new are what we are searching for.

For any doubts/queries, use our Discord server: <https://discord.gg/QMq6hsp3>.

Aryan Vora - 9769296849

Shreyansh Agarwal - 9044975252