

Precog Recruitment Task

Data

This is a database of ~80 million Indian district court data across states from the [Development Data Lab](#). Please download the data (.csv format) from [here](#).

To understand the tables, please refer to the metadata at the following links:

- [Act sections metadata](#)
- [All cases metadata](#)
- [Judges metadata](#)

Due to the large size of the data, you are free to use the following:

- Google Colaboratory or Kaggle workspace
- A subset of data by sampling rows from cases_[year].csv
 - For ex:
 - Rows for the time period 2010-2012 and 2022-2022.
 - Last 5/10 years

Analysis

1. Please do the necessary preprocessing and explore the data and come up with at least three insights about the data. Demonstrate your data exploration/analysis skills.
2. This analysis would be one of the main focus of our discussions in the interview. Surprise Us with your creativity!
3. Use plots to drive home the point of your insights.

Classification

1. Given the dataset - figure out a classification problem. Come up with an interesting classification problem and generate at least three insights.
2. Do all the necessary steps for building a classifier and generating insights. Following is the indicative (but not complete) list of steps that you should implement
 - a. Data splits
 - b. Trying out models
 - c. Reporting performance measures

Submission

- Put all your code/notebooks in a GitHub repository.
- Maintain a README.md explaining your codebase, the directory structure, commands to run your project, the dependency libraries used and the approach you followed.
- The link to the GitHub repository will be asked for during the interview.

Evaluation

It will be done based on the following:

- How do you handle and sample from large real-world datasets
- How interesting and creative your insights are
- How you present your code and findings

Please Note:

- You must attempt this task on your own.
- Please make sure you cite any external resources that you may use.
- We will also check your submission for plagiarism.

Bonus Task

For brownie points, you can read any one of the following papers.

1. [HLDC: Hindi Legal Documents Corpus](#)
2. [Effect of Popularity Shocks on User Behavior](#)

Based on your understanding, mention any three strengths and weaknesses of the paper along with any three improvements to the paper. Please fill out [this form](#) to report the above.

If you have any queries, please reach out to us at anmol.goel@research.iiit.ac.in.

All the best 😊