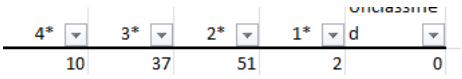
REF Project

Target

The target should be overall, where it will be predicting, whether it is 4*,3*, 2* and so on

Reason for choosing

- If we decide to predict the score for each Outputs, Impact or Environment, the performance metrics will be so low, due to the lack of data.
- If the model predict, the scores for each % of the submission meeting, then I can't assure the accuracy of such model
- I propose we get the highest % of the submission meeting, for example if overalls have the following



4*	3*	2*	1*	Unclassified
10	37	51	2	0

Since the highest is 2*(51), that will basically be the target

Inputs

There are 3 sets of Inputs I can suggest, they can be fused

Set 1

Main Panel,FTE of Submitted staff, % of eligible staff submitted

Set 2

ResearchDoctoralDegreesAwarded sheet (filter the unit of assessment name)

Number of doctoral degrees awarded in academic year 2013	Number of doctoral degrees awarded in academic year 2014	Number of doctoral degrees awarded in academic year 2015	Number of doctoral degrees awarded in academic year 2016	Number of doctoral degrees awarded in academic year 2017	Number of doctoral degrees awarded in academic year 2018	Number of doctoral degrees awarded in academic year 2019

- You can use the sum, get the average, std across years for each university

ResearchIncome sheet (filter the unit of assessment name)

Income source	Income for academic year 2013-14	Income for academic year 2014-15	Average income for academic years 2015-16 to 2019-20	Average income for academic years 2013-14 to 2019-20	Total income for academic years 2013-14 to 2019-20

- You can apply similar transformations to this also

ResearchIncomeInKind Sheet

Income source	Income for academic year 2013-14	Income for academic year 2014-15	Income for academic year 2015-16	Income for academic year 2016-17	Income for academic year 2017-18	Income for academic year 2018-19	Income for academic ye 2019-20

- You get the Idea

Set 3(Hardest)

Outputs sheet- The average Volume, Issue, First Page, Article Number, No of Additional authors), Number of articles for each university, Number of URLS each university has, Mode of open access status, citation applicable, Average citation count

ResearchGroups sheet - Total Number of research group

Performance Metric

- Accuracy

No of Rows/Data Points

- 364/4 = 91

Flow

- Prepare Data
- Split 60-40
- Explore data
- Transformation - Normalize numerical value, ordinally encode categorical values and the target
- Get correlation coefficient
- Train Data
- Evaluate with test
- Create Pipeline
- Train pipeline on whole data
- Save Pipeline