**Accident Risk Index**

The rise in vehicles on the road will also lead to multiple challenges and the road will be more vulnerable to accidents. Increased accident rates also leads to more insurance claims and payouts rise for insurance companies.

In order to pre-emptively plan for the losses, the insurance firms leverage accident data to understand the risk across the geographical units e.g. Postal code/district etc.

You have to predict the “Accident\_Risk\_Index” against the postcodes. Accident\_Risk\_Index (mean casualties at a postcode) = sum(Number\_of\_casualities)/count(Accident\_ID).

**Evaluation:**

Split the dataset into 75%/25% (train-test) using stratified shuffle split and for training and testing, use metric ‘[Root Mean Square Error](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html)’.

Working example:

|  |  |  |
| --- | --- | --- |
| Train Data (given) |  |  |
| Accident\_ID | Postcode | Number\_of\_casualities |
| 1 | AL1 1JJ | 2 |
| 2 | AL1 1JP | 3 |
| 3 | AL1 3PS | 2 |
| 4 | AL1 3PS | 1 |
| 5 | AL1 3PS | 1 |

**Features in file ‘dataset.csv’:**

1. 'Accident\_ID',
2. 'Police\_Force',
3. 'Number\_of\_Vehicles',
4. 'Number\_of\_Casualties',
5. 'Date',
6. 'Day\_of\_Week',
7. 'Time',
8. ‘Local\_Authority\_(District)',
9. 'Local\_Authority\_(Highway)',
10. '1st\_Road\_Class',
11. '1st\_Road\_Number',
12. 'Road\_Type',
13. 'Speed\_limit',
14. '2nd\_Road\_Class',
15. '2nd\_Road\_Number',
16. 'Pedestrian\_Crossing-Human\_Control',
17. 'Pedestrian\_Crossing-Physical\_Facilities',
18. 'Light\_Conditions',
19. ‘'Weather\_Conditions',
20. 'Road\_Surface\_Conditions',
21. 'Special\_Conditions\_at\_Site',
22. 'Carriageway\_Hazards',
23. 'Urban\_or\_Rural\_Area',
24. 'Did\_Police\_Officer\_Attend\_Scene\_of\_Accident',
25. 'state',
26. 'postcode',
27. 'country'