**YELP CHALLENGE: Determining Future Health Score for Businesses**

Stakeholders: Yelp, Government

* **What you plan to do?**

We plan to predict health score for restaurants in the USA

* **Why it is important?**

It will be useful for the Government to ensure Public Safety and further allocate resources in the best possible way based on the health score.

It will be useful for Yelp to collaborate with Businesses to provide better service from their side.

1. For government - Prediction of a standard health score for each restaurant (based on available feature data) can help the government to decide the risk of a restaurant. This can help them to allocate limited official resources for restaurant inspection efficiently. Hence it will ensure public health and safety.

2. For yelp - (i) They can collaborate with businesses to provide recommendations to improve their business and user needs based on standard health score. This can provide mutual benefit to both yelp and businesses. (ii) They can collaborate with government to provide health score and health analysis provided by this solution.

* **What algorithms do you plan to employ?**

|  |  |
| --- | --- |
| **Components** | **DM algorithms used** |
| 1. To predict health score (output class) based on input features (such as ratings, reviews, violations, and others): The data for features would be taken from the yelp dataset and online open source data provided by government health agencies. The past data for health score would be collected from an open source data provided by government health agencies. | Supervised Machine learning |
| 2. To fill missing entries in yelp data: There are a lot of features that are available for some restaurants and not for others. Multiple restaurants can be compared to find similarity in restaurants. Multiple features can be compared to derive similarity in features. | Find similar items and users, recommendation systems |
| 3. Clustering of restaurants based on area [Extended goal if time permits]: To find which areas have good restaurants and which ones have poor restaurants. This can be useful for the government agencies when they want to see areas performing good vs bad for restaurant inspection and public safety. This can also be useful for users when they want to search for areas with good restaurants. | Clustering techniques (such as k-means) |

(i) Supervised ML techniques

(ii) Find Similar Items, Recommendation (Fill missing entries)

Further Extend if time permits: Clustering based on Area

* **How you plan to evaluate?**

Precision

Recall

* Possible Datasets to use: Yelp Dataset, Data provided by country government(Health)

\*\* area based analysis to find good/bad restaurants for area

**Team**

Ajay Anand : ajayanan@usc.edu

Devershi Purohit : dupurohi@usc.edu

Rajdeep Kaur : kaurr@usc.edu

Rimsha Goomer : goomer@usc.edu

Things to look :

* Which states/area data is available . for example LA County :
* Yelp dataset restaurant intersect with data from above

LA County Sample Data:

<https://data.lacounty.gov/browse?category=Health&utf8=%E2%9C%93>

LA County Data:

Activity Date,

Id,

restaurant name/address,

code description(Ex: row risk)

Violation Id: Ex: bad food

Health Score (0-100)

Grade (A,B etc.)

Yelp Data

1. Restaurant name
2. Rating
3. Review
4. Price
5. Phone no.
6. Takes reservation
7. Delivery option
8. Take out
9. Accepts credit card/apple pay
10. Good for (Breakfast,Lunch,Dinner)
11. Good for Kids
12. Good for groups
13. Attire : Casual
14. Ambience
15. Noise Level (Loud, Average)