DETERMINING A FEATURE VECTOR FOR CLASSIFICATION THE FOOD SERVING VENUE ACCORDING TO RISK FOR A PUBLIC HEALTH

NIHAD HASIC

THE PROBLEM

- Sanitary inspection of various cities in United States publish the results of checking the sanitary and other relevant conditions of venues that serve a food in the form of dataset with list of venues' metadata and the result of the inspection (ranking by the risk) on the healthdata.gov web site (http://www.healthdata.gov). On the other side, the Foursquare dataset contains an attribute "likes" that contains the number of times the users gave a like (prefer, vote up) that restaurant among others and attribute "rating" which represents the average of all ratings given by users. It is interesting to see how objective the users in their likes and ratings are/could be from the perspective of sanitary conditions, or how much are sanitary conditions relevant to users (guests) in their decision to like or not to like or how to rate a restaurant.

THE IDEA

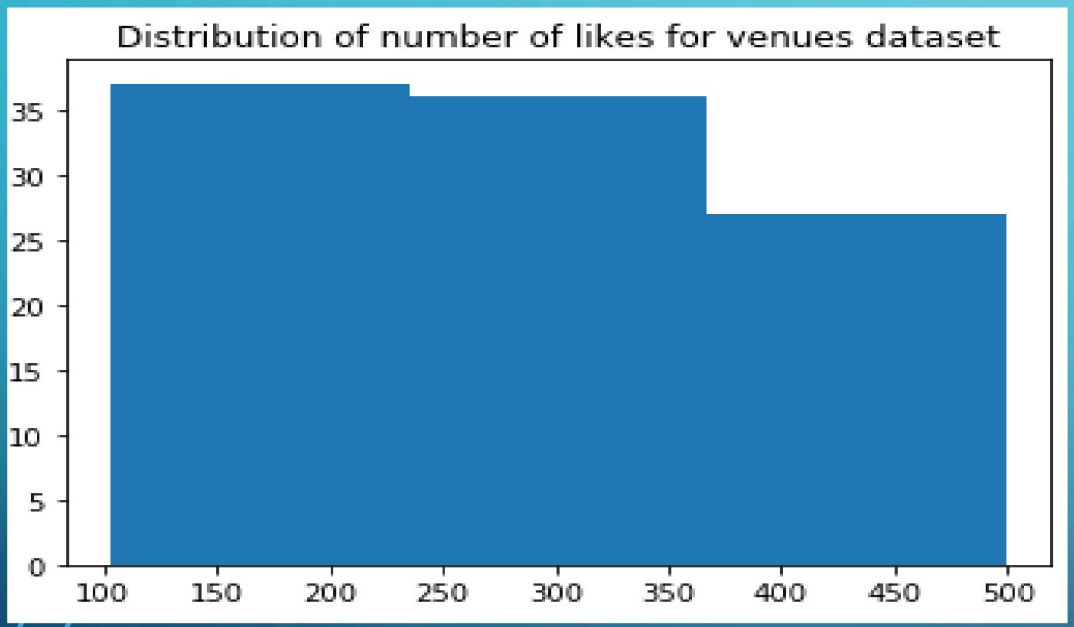
- Idea of the project is to correlate this dataset with Foursquare database for one city (Chicago), and check if attributes “likes” and “rating” can be a good predictor for a risk category of a restaurant according to the inspection results . Based on data for the reference city (Chicago) determine a set of features relevant to classify the restaurant in another city as risky or another class, according to available list of risk classes in the results of inspections (multivalue classification). As evaluation dataset will be used the exact results of inspection control for the second city (San Francisco), acquired from healthdata.gov.

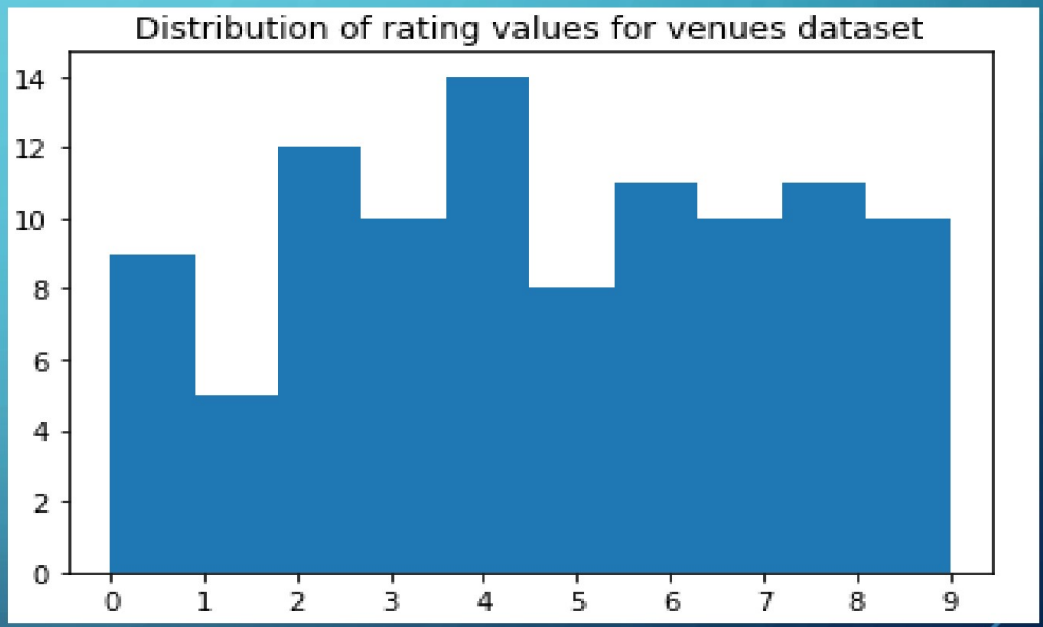
TRAINING SET

- Matching Foursquare set with 100 venues for Chicago and inspection control dataset with 195000 rows
- Matching according to minimal value of Levenshtein's distance between name and address of venue

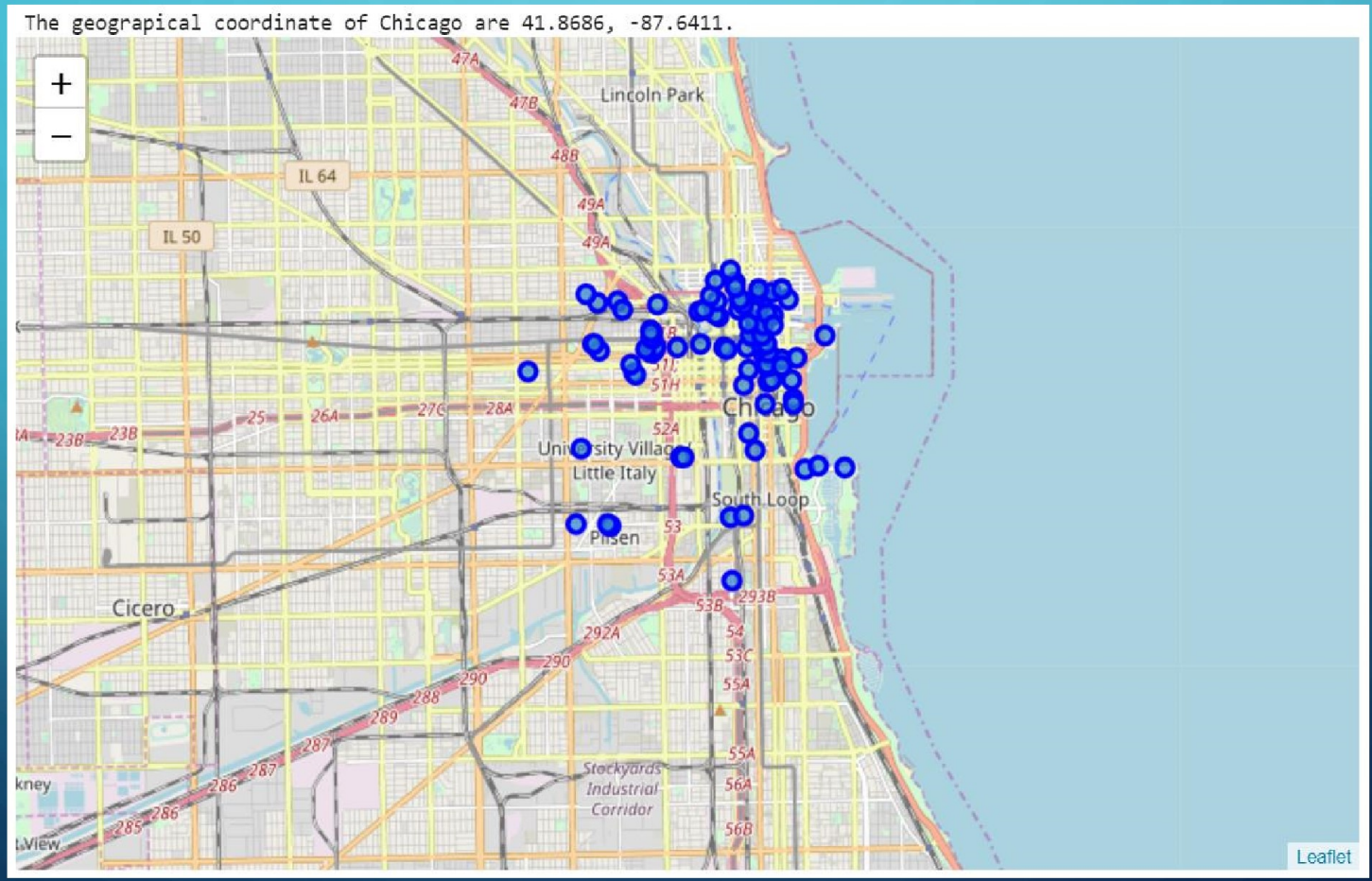
LEVENSHTEIN'S DISTANCE

TRAINING DATASET

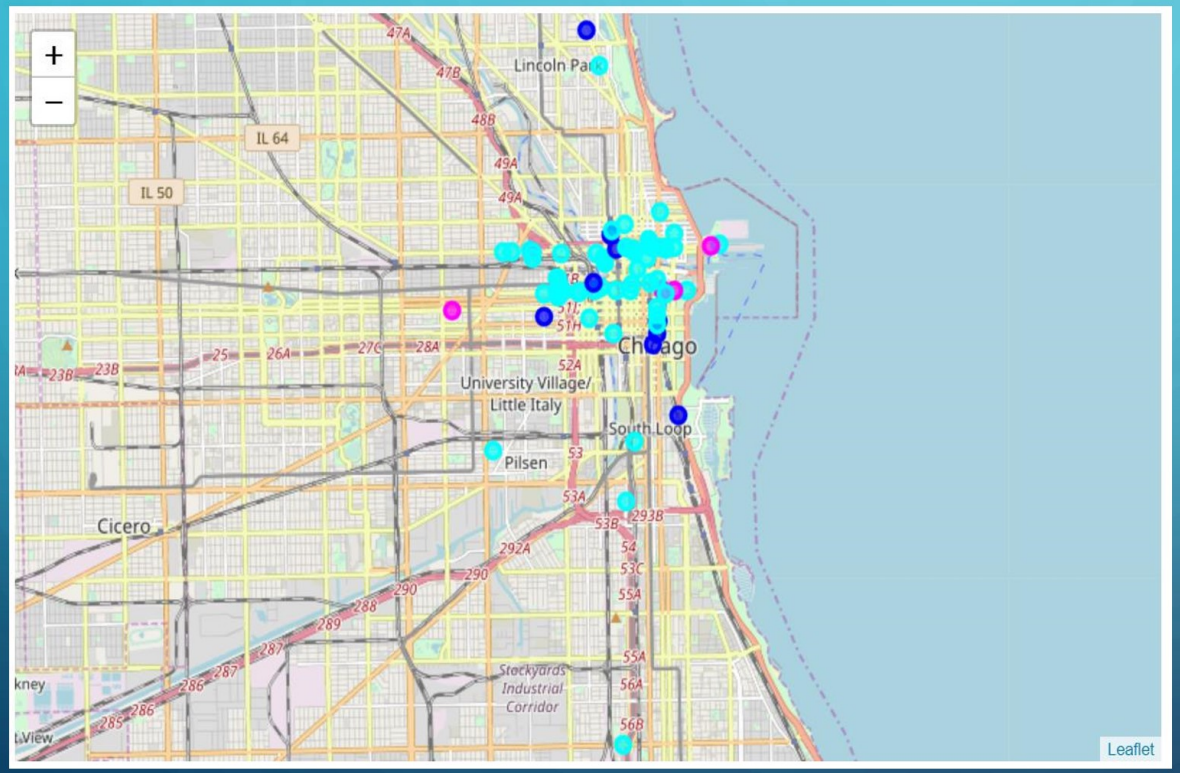




TRAINING DATASET - SPATIAL



TRAINING DATASET - MERGED

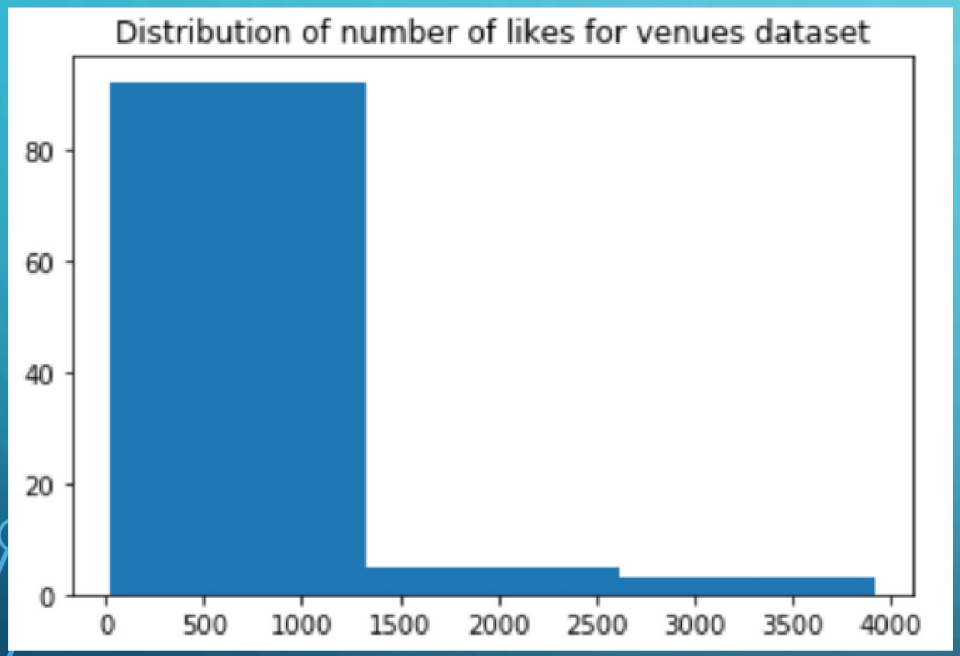


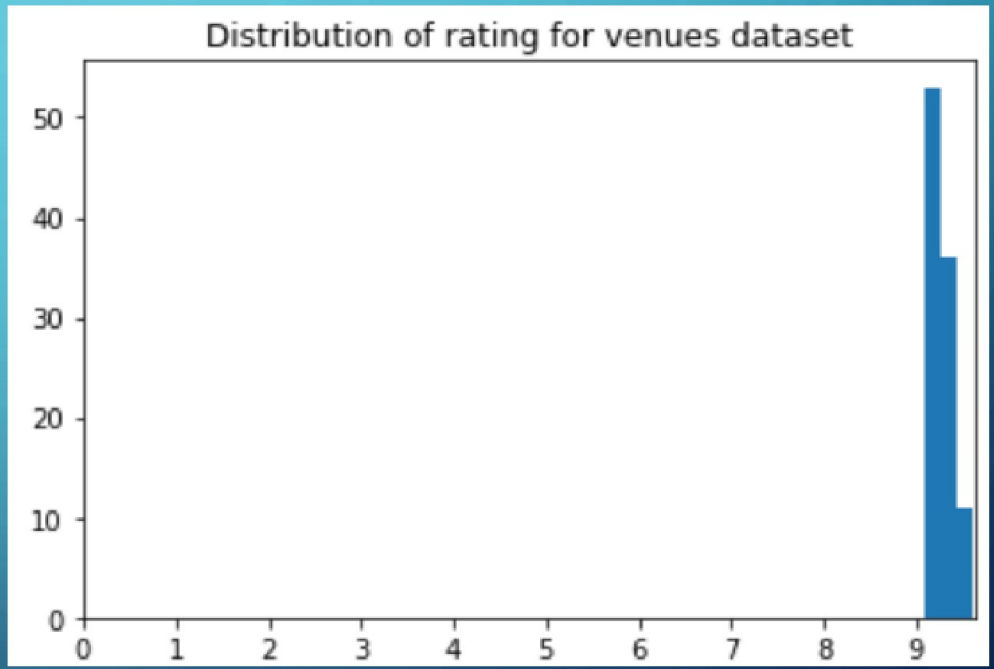
Color	Coded risk value	Risk
cyan	1	High
magenta	2	Moderate
blue	3	Low

TEST DATASET

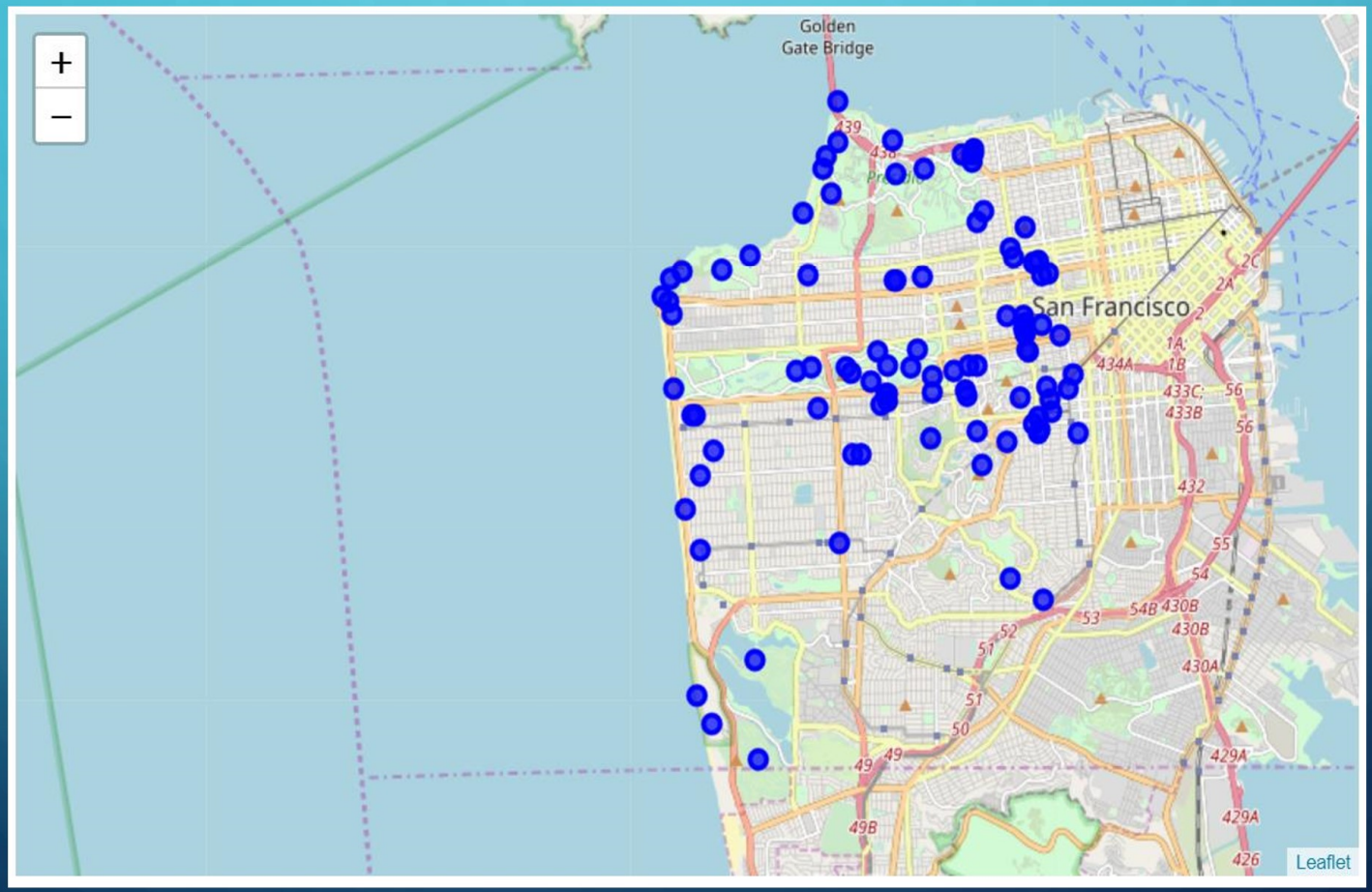
- Matching Foursquare set with 100 venues for San Francisco and inspection control dataset with 54000 rows
- Matching according to minimal value of Levenshtein's distance between name and address of venue

TEST DATASET

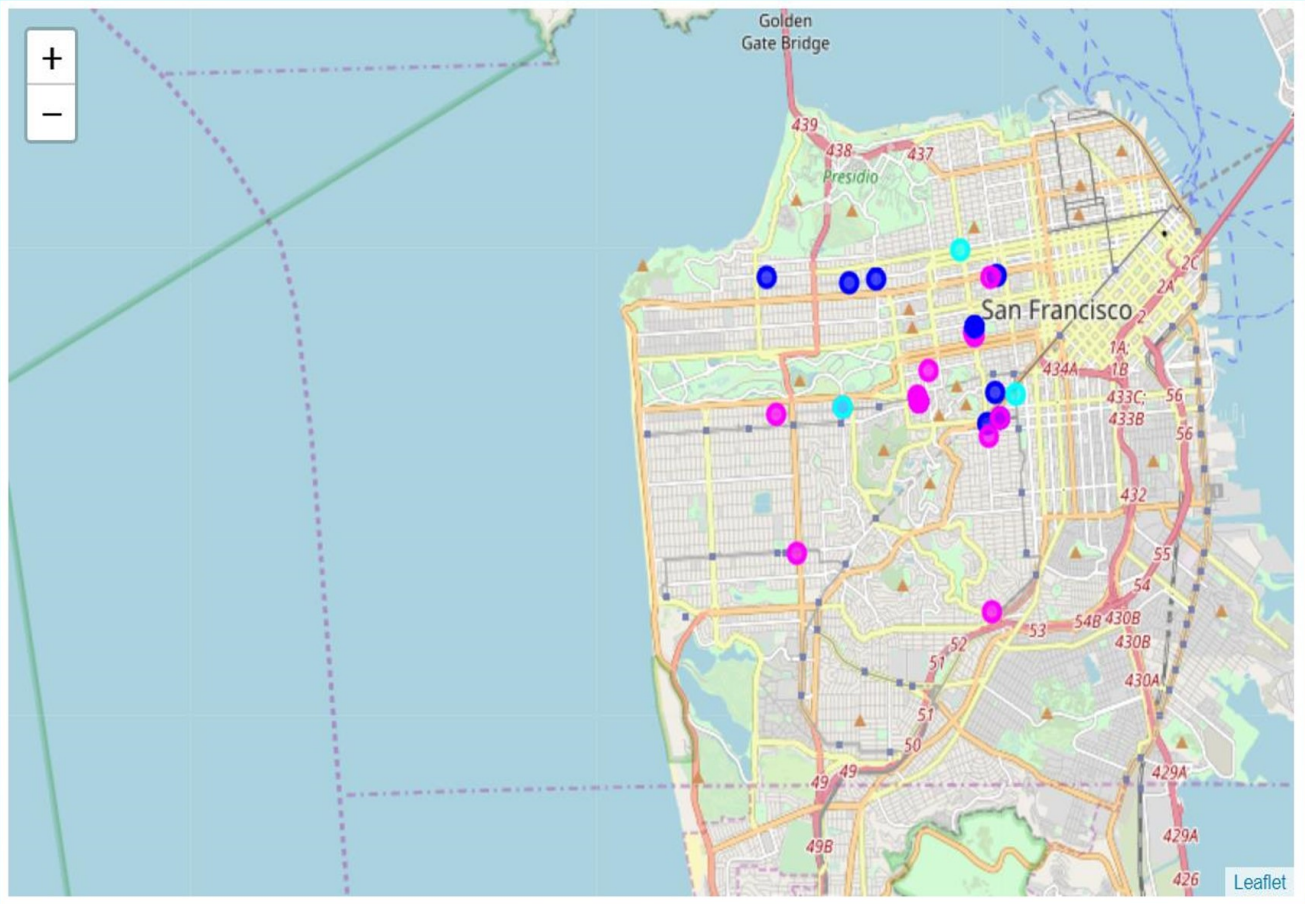




TEST DATASET - SPATIAL



TEST DATASET - MERGED



Color	Coded risk value	Risk
cyan	1	High
magenta	2	Moderate
blue	3	Low

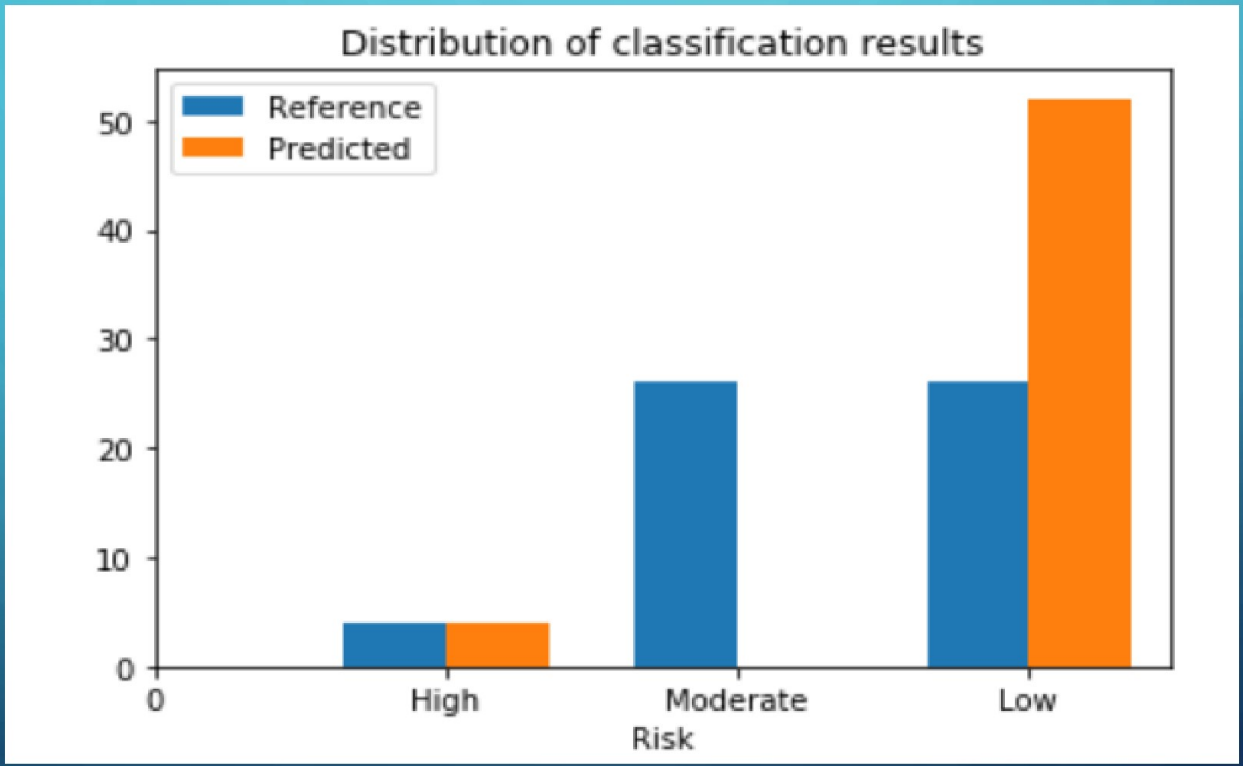
CLASSIFICATION

- 3 classifiers evaluated:
  - K-nearest neighbours
  - Logistic regression
  - Decision trees

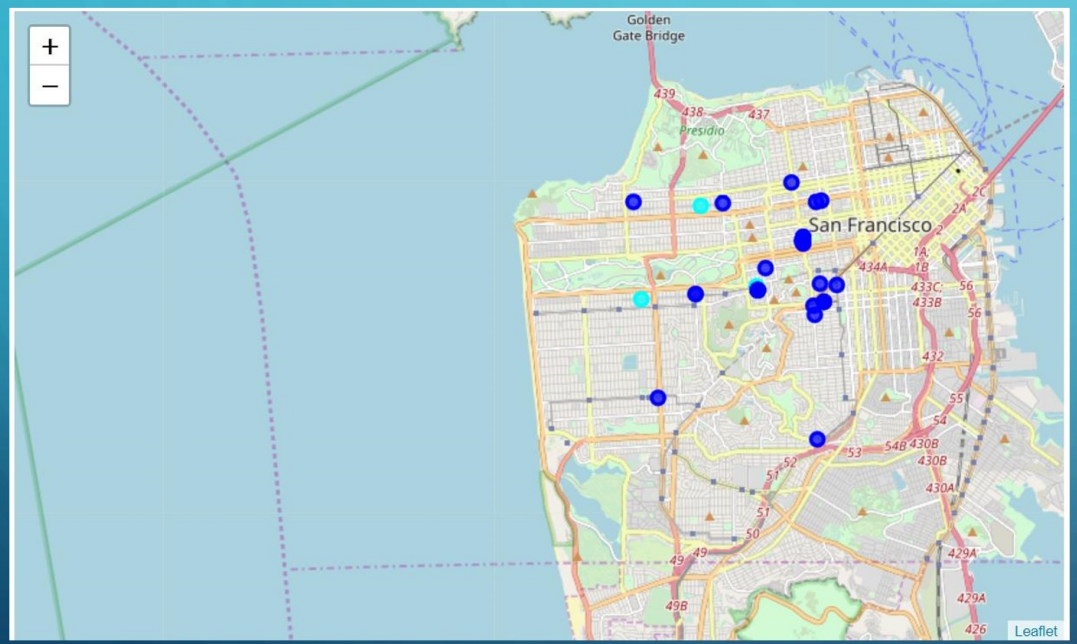
- Classification metrics:

Algorithm	Jaccard similarity index	F1 score
Logistic regression	0.44643	0.29762
KNN (K=4)	0.08929	0.04424
Decision trees (max depth = 6)	0.07143	0.00969

RESULTS - LOGISTIC REGRESSION

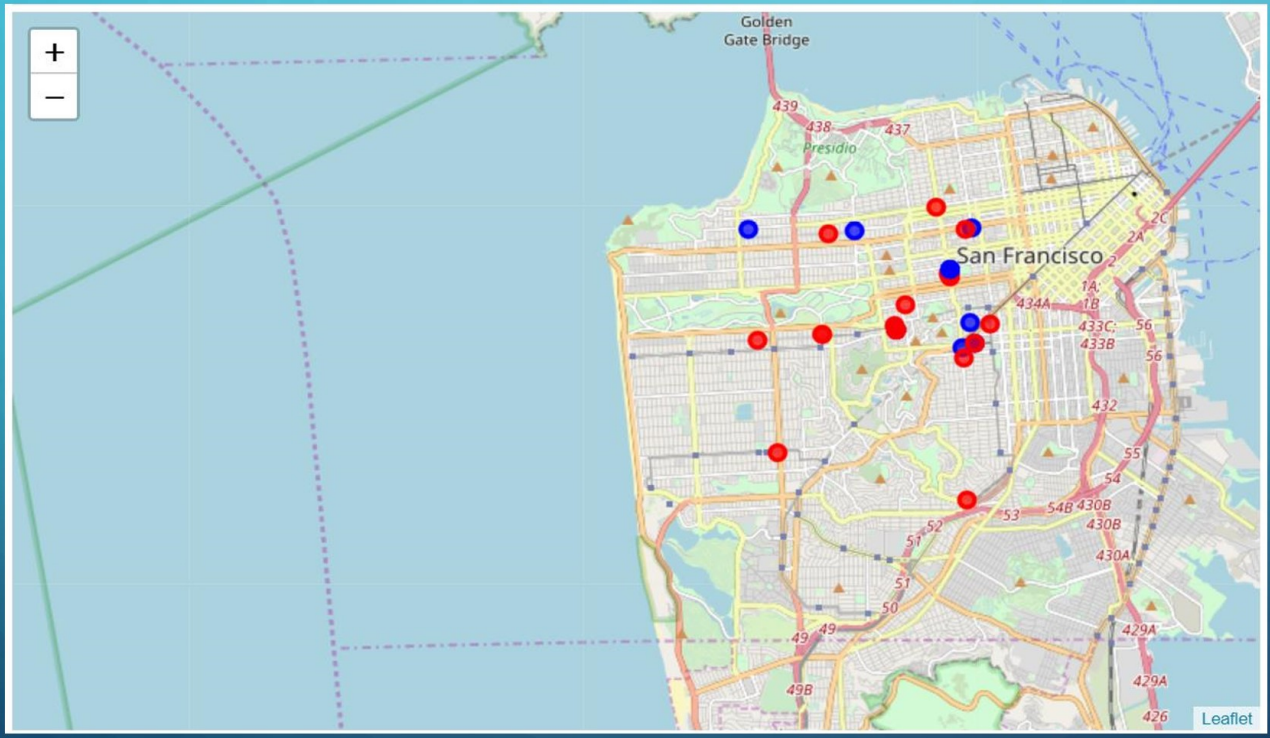


RESULTS - SPATIAL



Color	Coded risk value	Risk
cyan	1	High
magenta	2	Moderate
blue	3	Low

RESULTS - SPATIAL



CONCLUSION

- In this analysis it couldn't be proven that number of likes and rating on the Foursquare could be good predictors of sanitary conditions or results of inspection controls. Value of Jaccard's similarity index of 0.44643 in the case of logistic regression tells that there is a possibility with the bigger training sample to prove a relationship between variables and significance of number of likes and rating as the features in the classification of venues according to sanitary risk.