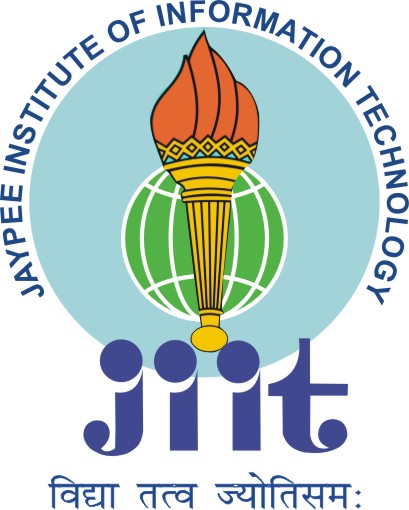
**MINOR PROJECT (CSE) REPORT**



**SUBMITTED TO: SUBMITTED BY:**

**PROF. YAMUNA PRASAD ISHANK AGGARWAL (15103041)**

**PROF. SHIKHA MEHTA SARTHAK JOSHI (15103102)**

**OSHEEN SINGH (15103189)**

**UDIT GOEL (15103376)**

1. Mention your individual contribution in below mentioned table. Out of 100 percentage, how much work has been done by individual, total should be 100. **(Form Should be FrontPage of the report and all below** **mentioned content is part of Minor Report)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Project Title: Cuisine Analyser** | | | |
| **S. No.** | **Enrollment No.** | **Student Name** | **Work Done (in percentage) out of 100 contribution of each member in group** |
| 1. | **15103041** | **Ishank Aggarwal** | **25%** |
| **2.** | **15103102** | **Sarthak Joshi** | **25%** |
| **3.** | **15103189** | **Osheen Singh** | **25%** |
| **4.** | **15103376** | **Udit Goel** | **25%** |

1. ***A. EXISTING GITHUB LIBRARY (5 Marks)***  
   1. Git hub URL including complete author details

Github URL:- https://github.com/conwayyao

Author Details:- <https://www.linkedin.com/in/catchconway>

Currently working in FiscalNote as a Data Product Manger, Conway Yao has majored from Georgetown University in BSFS, Science, Technology, and International Affairs

2. Demonstrate executable existing Git hub project (chosen one)

Existing Project Repository URL:- https://github.com/conwayyao/Recipe-Analysis

3. Number of Issues mentioned in the project and separately state the issues which you  
would like to resolve(if any)

Issues Mentioned in the Project:

* Accuracy of regression and naïve bayes is low
* deciding on how to use the two scoring algorithms- either take the mean of both or find the ones that achieve high score on one but not the other and classify them separately
* The bag of models is inelegent.There should exist a method to use the ingredients as tokens for a model rather than deconnect back to a bag of words (Ingredient tokenization).
* Anything that can be done to improve the situation of multi-cuisine designations to improve the accuracy. (Multiple Cuisine designation).
* Clustering analysis- k-means, tfidf, etc
* Pricing analysis if we can find an API that accepts ingredients and returns price per serving.

Issued which we would like to resolve:

* Propose a recipe given set of ingredients
* KNN classifier using word count and tf-idf to recommend similar recipes.
* to find relationship between popularity(rating) vs ingredient count, , rating and cooking time, cuisine label and recipe label.
* to find relationship between cooking time vs cuisine(people of which country(eg. thai) tend to spend more time in cooking)
* to predict the histograms for relationship between precision, recall and f1 vs different classifiers we tried out
* to predict the histograms for relationship between accuracy vs different classfiers we tried out
* to depict the confusion matrix plot of different classifiers.
* code reimplementation on graphlab
* using recipe instructions as parameter in order to compare cooking techniques and methods
* unsupervised NLP approach to select dish types from recipe names would be challenging as many recipe names are confounded by adjectives and descriptors or do not use standard english terms (eg. “Garllic Aoilli-Dipped-Salmon ceviche”- Is this salmon recipe or ceviche recipe? . we would be working on solving this ambiguity.
* to infer the cuisine based on generalized characteristics. This will be little tricky as its very easy to find recipes with exact/similar ingredient matches.
* fitting multidimensional regression curve to the scatter plot between ingredients and cooking times.
* Clustering analysis- k-means, tfidf, etc
* Publishing interesting results or infographic online using static images or Bokeh/D3
* Is there a relationship b/w recipe rating and recipe costs?
* What other relationships exists in the data?- plotting graphs, histograms, etc.
* Which recipes can I make with a few more ingredients?
* Given a set of ingredients, which cuisine/recipe is it most like?
* To prepare confusion matrices of naive bayes, logistic regression and random forest.

4. Existing code files with description of file content (functionality performed)

|  |  |
| --- | --- |
| Code file name ***(EXISTING)*** | Functionality DESCRIPTION |
| cuisineAnalyzer.py | Explores the patterns existing among different recipes of various cuisines |

***B. Reusing of content (2 Marks)***  
5. Problems faced while reusing existing code:

|  |  |  |
| --- | --- | --- |
| Issue description | Issue location(code file name) | Resolved / not resolved |
| * Missing parameters * Missing codes | cuisineAnalyzer.py | Resolved |

***C. Current Project***  
•Problem Statement [150 Words] ***[ 3 Marks]***

Over the years, people have tried to explore new ingredients and incorportate them into recipes or produce new recipes all together. In this research project, we explore the patterns related to the ingredients of a cuisine.

Recipes tend to encourage a slavish devotion to the recipe and divert attention from the more important part of cooking, the physical abilities (or, in my case, the lack thereof) of the chef. Foodies tend to privilege the provenance of obscure ingredients ("coulis of feather saffron hand-picked from a seaside village in Morocco").

We are curious how much variation exists between dishes, and whether such variation is warranted. Are there really 5000 ways to cook a steak, or are many of these variations superfluous? We hope to use a data-science approach to see if the cooking wisdom of the crowds have arrived at the same answers, and if they match those of traditional experts.

To answer questions like most common ingredients for a particular cuisine and most unique recipe for each cuisine based on its ingredients, this repository was initiated and we have chosen this repository to analyse the performance of the existing code present there and to implement our advancements as a part of our research project.

• Software / technologies and databases used for implementation and connectivity

* + - Jupyter Notebook (Numpy, Pandas, scikitlearn, matplotlib, etc. Included)
    - Graphlab
    - Anaconda

• List number of functionalists added in existing code in following format. ***[ 10 Marks]***

|  |  |
| --- | --- |
| Code file name ***(CURRENT)*** | Functionality added |
| cuisineAnalyzer.py | * propose a recipe given a set of ingredients - this question has been answered * used k-nearest neighbours (KNN) classifier to recommend similar recipes i.e. to answer question like given a recipe, which recipes are most similar to it?. This feature involves building nearest nearest neighbour model by using both word count and TF-IDF * In order to improve accuracy of regression and naive bayes, we tried 3 more classifiers- Random Forest, K-cross validation and KNN. * Relationship between cooking time vs cuisine is analysed. (for eg. people of which country (eg. thai) tend to spend more time in cooking.), * Relationship b/w rating and cooking time, rating and ingredient count, cuisine label and recipe label have been observed. |

• Novelty of added functionality (1-2 functionalists different than existing similar sort of projects) ***[2 Marks]***

*(i) adding features*

*-> Prediction of cuisines based on sentiment analysis*

*-> Recommending similar cuisines to the predicted cuisine(in case u did not like it)*

*-> if we want a particular cuisine & we have 1-2 missing ingredients to make it …what other cusines can we make*

*-> what cuisines can we make from a few more ingredients.*

*(ii) improving prediction accuracy.*• Draw Functionality Mapping diagram of complete project ***[3 MARKS]***

ML Model

Training Data

Feature Extraction

**X ŷ (Prediction)**

**y ŵ (update)**

ML Algo

(

Quality Metric

**Ŷ compared with y**

• Demonstration of executable project (at-least 70-75% code should be complete by mid  
phase) ***[20 Marks]***