**ChefTime**

Mohammad Mohsin Hussain (2019255)

Rohit Roy (2018259 )

Ashmeet Singh (2019412)

Note: We have added in detailed code documentation and plots in the .ipynb files

**Objective**

We aim to predict preparation time for recipes using:

* + Ingredients, :- these are given in form of sentences
  + Instructions :- these are given in form of paragraphs

The preparation time of a recipe is mainly dependent on these variables as it refers to the time we are busy in preparation for food in our kitchen, mixing, mashing, stirring the ingredients, etc, whatever the recipe tells us to do.Along with this,we also do exploratory analysis and show plots which shows general trend in some features such as cuisine, Diet and courses chosen by people.

This project would be helpful to predicting preparation time from ingredients and instructions in sentences form. This idea is specifically helpful in case we want to know the estimated time that a particular food item would take in restaurants.

**Dataset**

We are using Kaggle dataset for our project.

Link : <https://www.kaggle.com/datasets/kanishk307/6000-indian-food-recipes-dataset>

**About dataset**

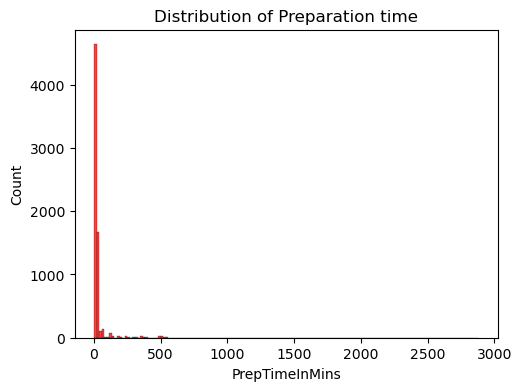
The dataset has 6000+ recipes data fetched from <https://www.archanaskitchen.com/>

It has around 15 columns,we have focused on the following 6 columns namely:

* + TranslatedIngredients
  + TranslatedInstructions
  + PrepTime
  + Course
  + Cuisine
  + Diet

**Analysis**

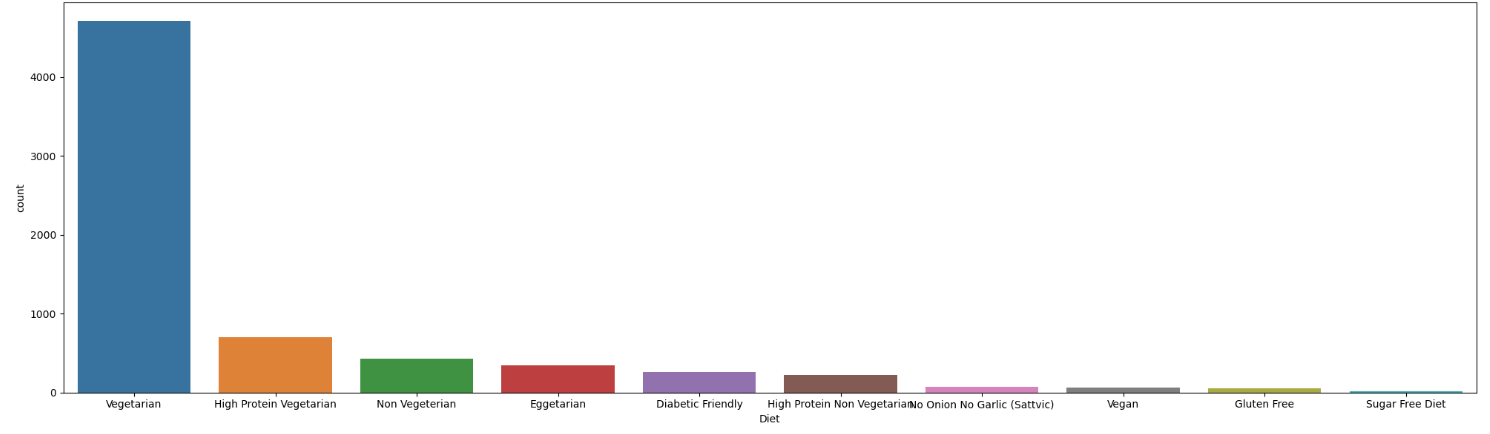
We first analyzed our data set to get some inferences about it.

**Frequency-Plot of preparation time**

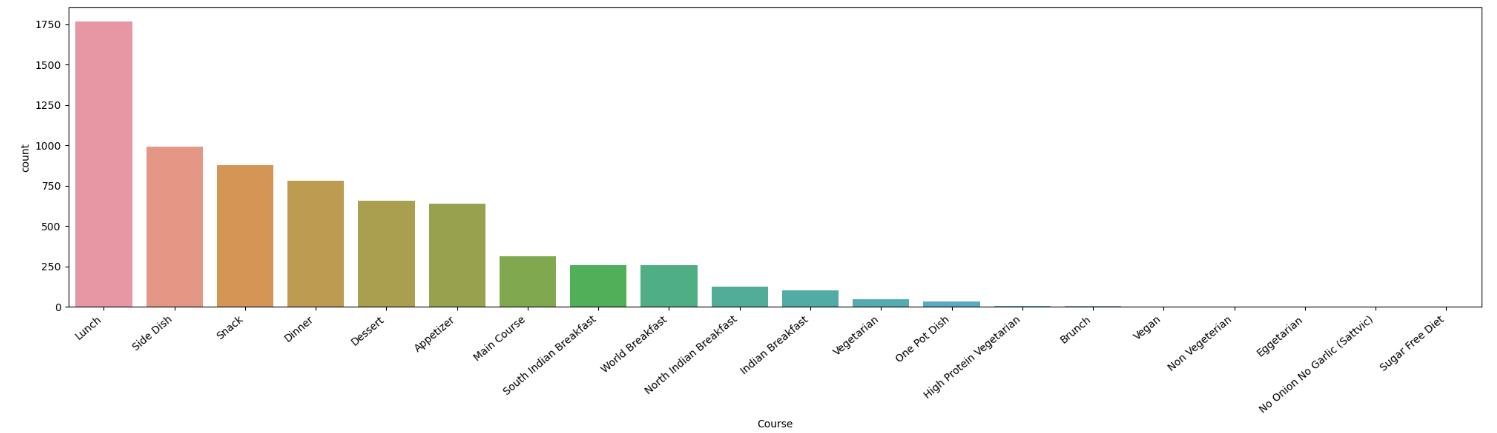
We can clearly notice from the bar graph that preparation time in data is highly biased.

From this plot, we realized that most data are concentrated at lower values of preparation time.

**Frequency-Plot of Diet.**

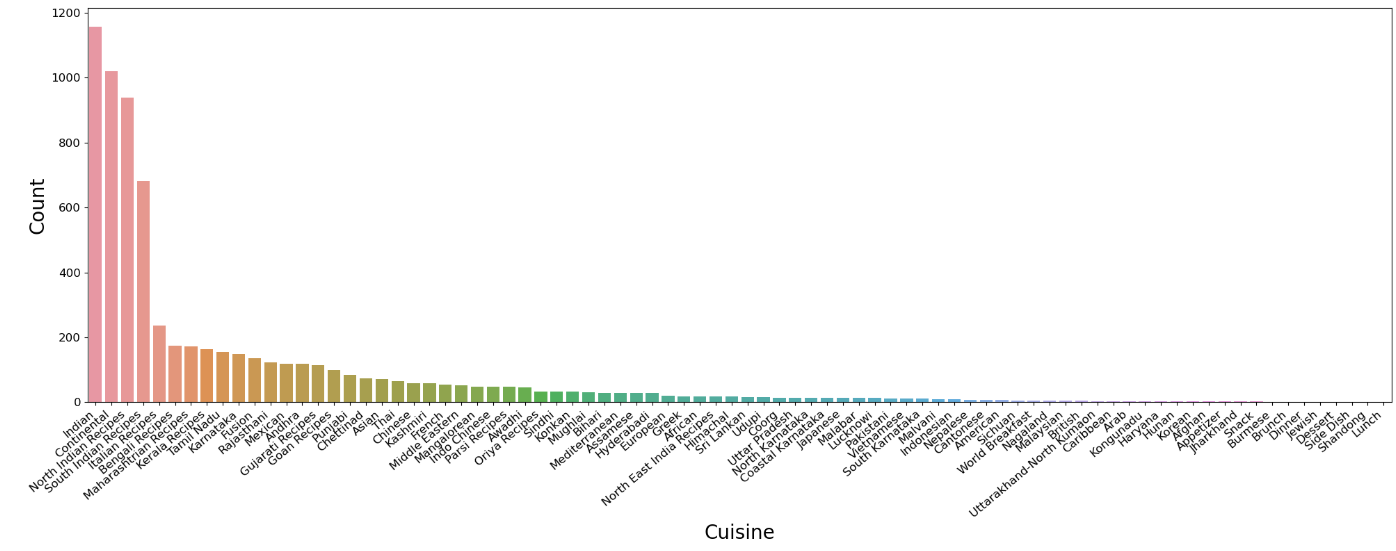
****From the sorted frequency plot we can see that Vegetarian is highest and rest are few.

**Frequency-Plot of preparation Course.**

****

From the sorted frequency plot we can see that many lunch side dish and snacks dominate the data.

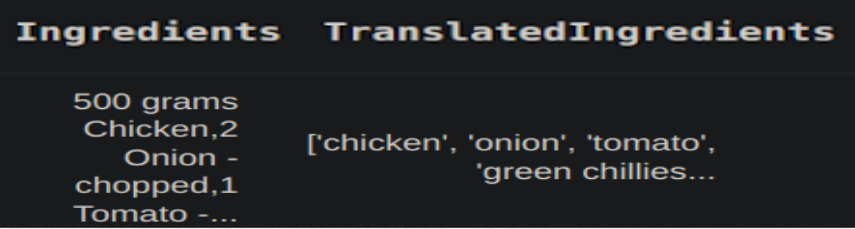
**Frequency-Plot of preparation Cuisine.**

****

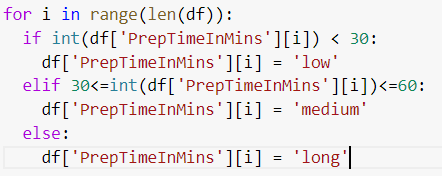
We can see that most cuisine data is labeled as Indian followed by continental then specific Indian states with some mixture of cuisine from other countries.

**Evaluation**

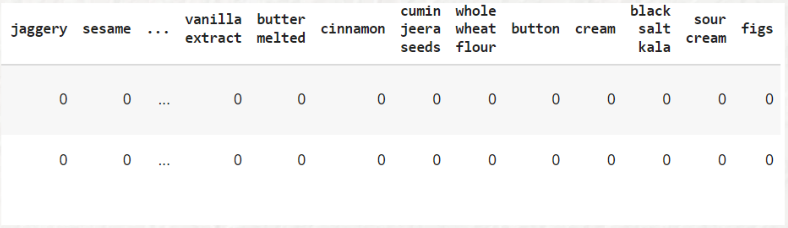
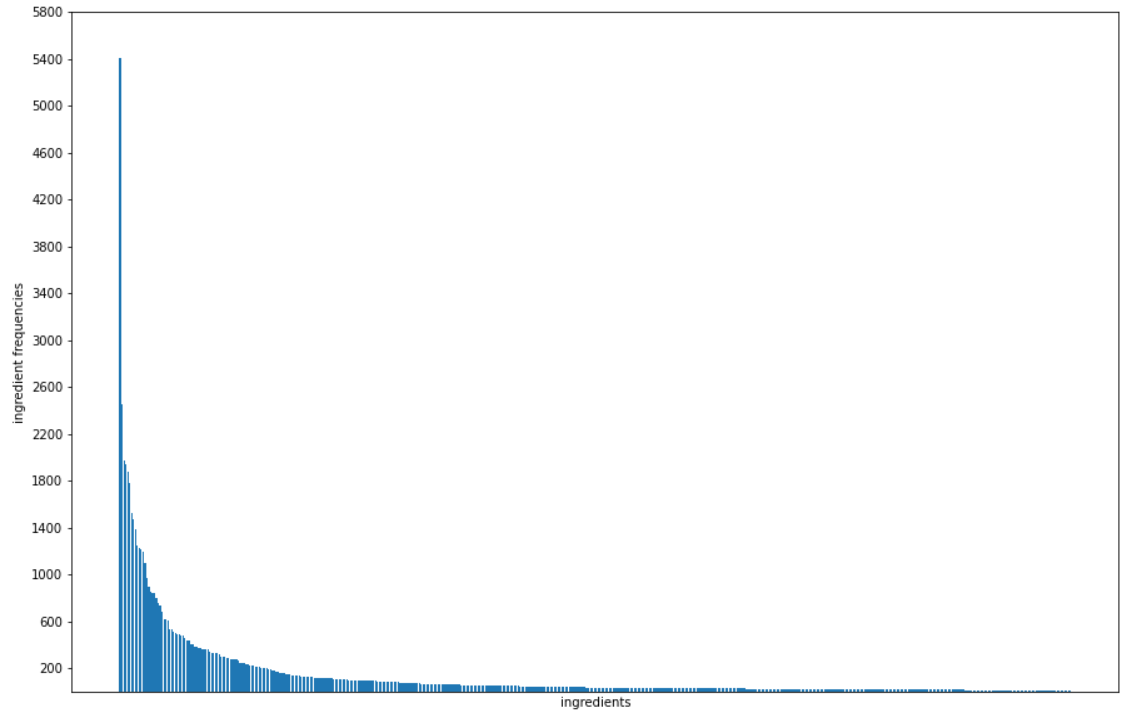
**Preprocess ingredients**

First, we preprocessed the ingredients. We used an ingredient parser library(pying) to extract ingredient names from the ingredient column. We removed the rows where nan values were present. Removed those rows where hindi letters were present in the TranslatedIngredients column. Then we Removed those rows where hindi letters were present in the TranslatedIngredients column.

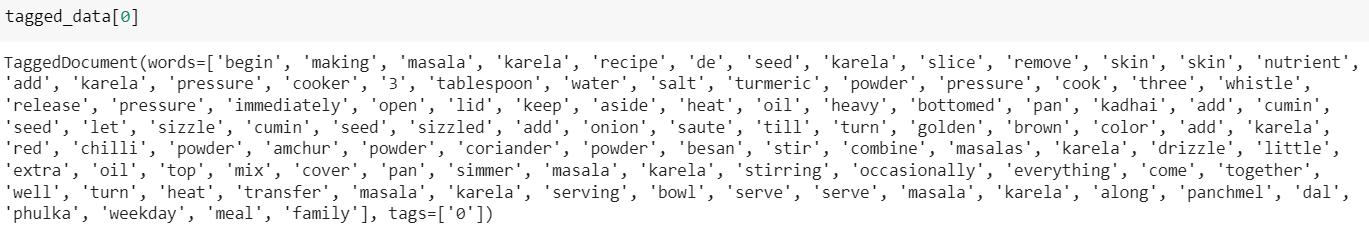
**Convert preptime into categories**

Since the task is tedious, we converted preptime into categories. We converted PrepTime in three categories : Low(<30 minutes), Medium(between 30-60 minutes) and Long (>60minutes). 

**Select top ingredients for one-hot encoded features**

After checking the frequencies of ingredients, we saw that a lot of ingredients were irrelevant as their frequencies were very low. So we took the top 300 ingredients from each of the ‘low’, ‘medium’ and ‘long’ classes. Then we one-hot encoded those ingredients as columns in the dataframe.

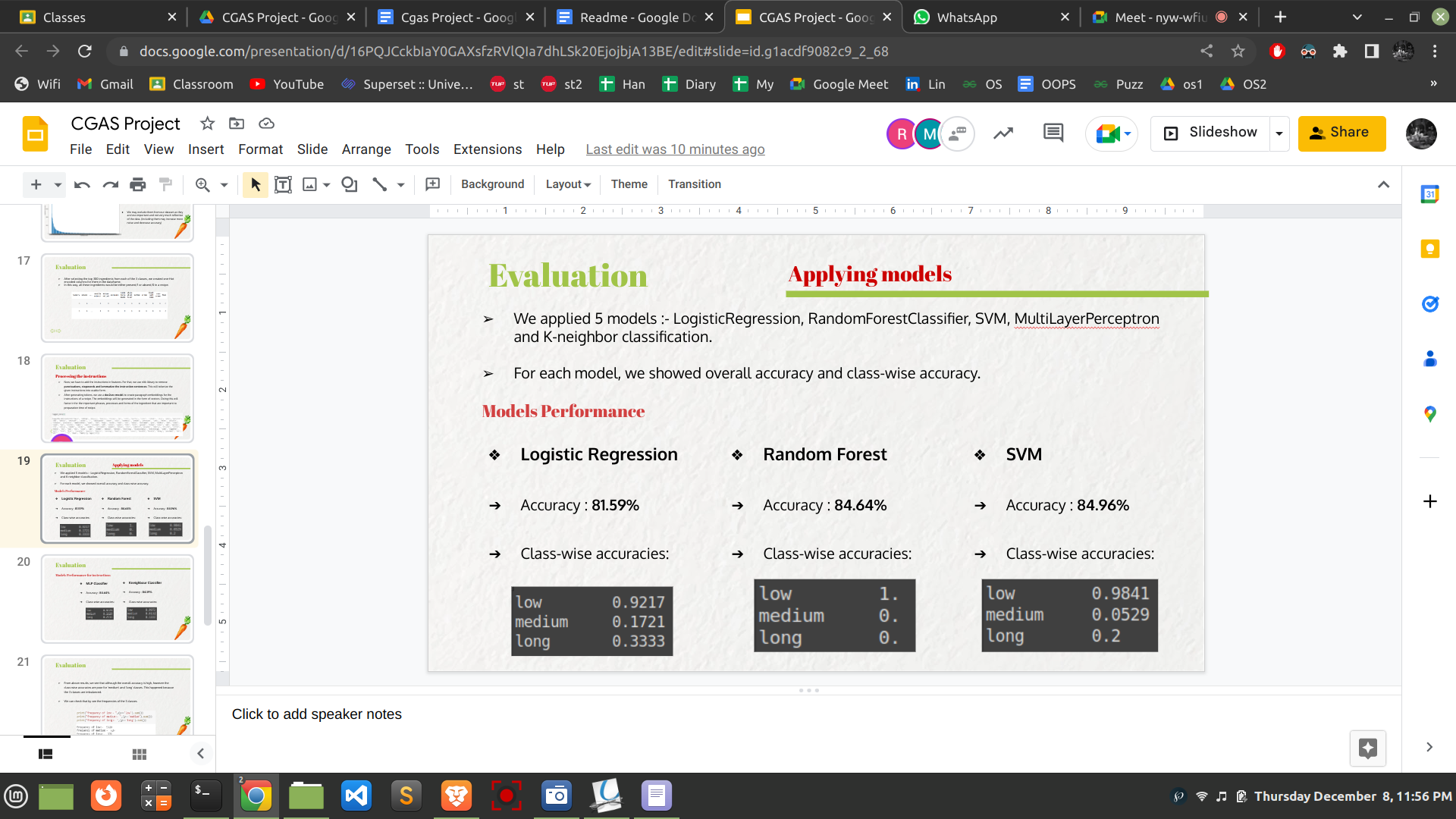
**Preprocessing the Instructions**

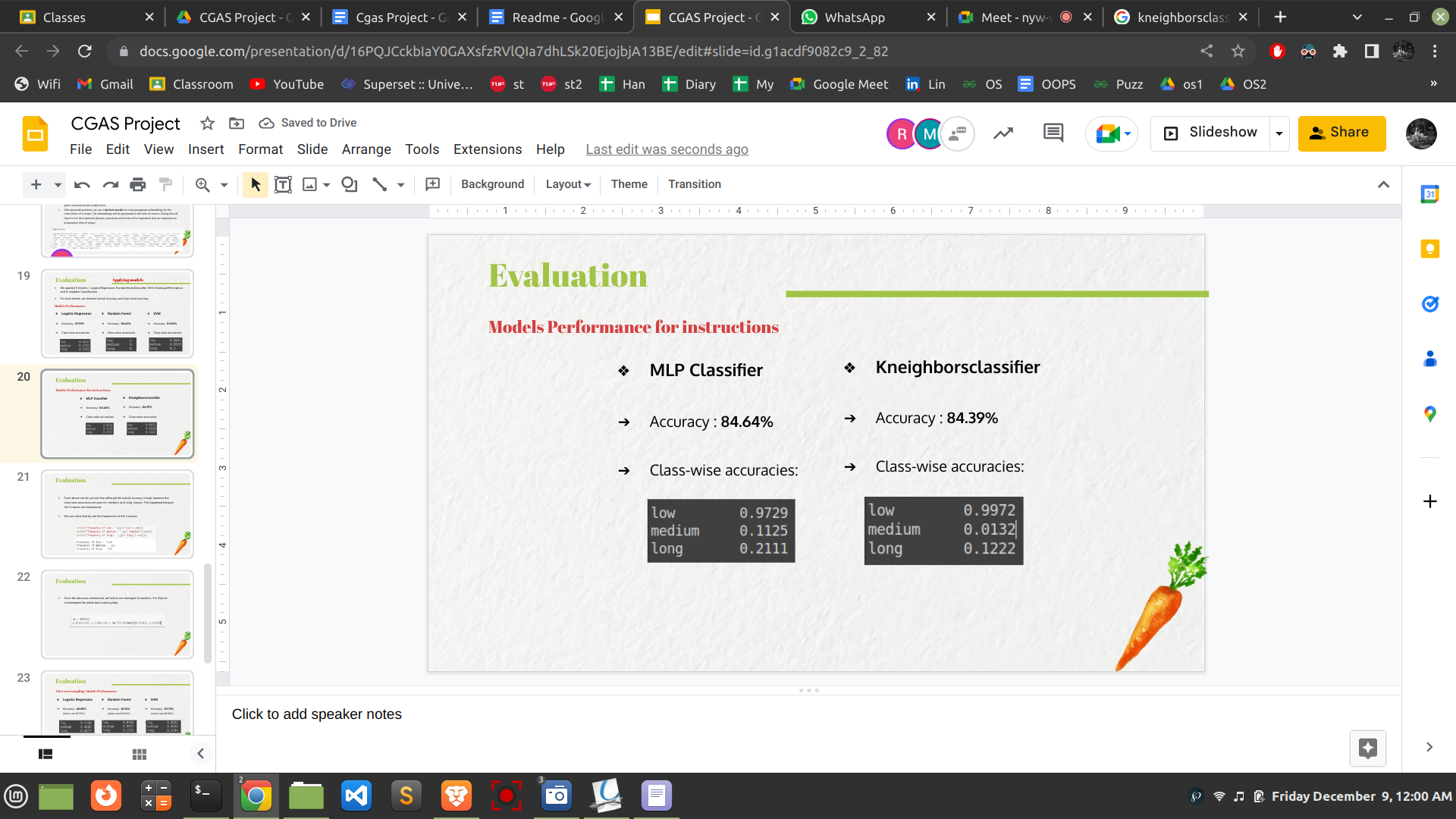
The instructions had many unwanted characters,in order to obtain better results we have to preprocess our data. First, we have to add the instructions in features. For that, we use nltk library to remove **punctuations, stopwords and lemmatize the instruction sentences.** This will tokenize the given instructions into usable form.After generating tokens, we use a **doc2vec model** to create paragraph embeddings for the instructions of a recipe. The embeddings will be generated in the form of vectors. Doing this will factor in the important phrases, processes and forms of the ingredient that are important to preparation time of recipe like this.

**Applying Models**

We applied 5 models :- LogisticRegression, RandomForestClassifier, SVM, MultiLayerPerceptron and K-neighbor classification.

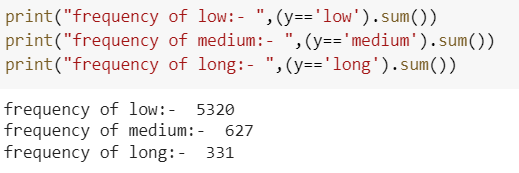
For each model, here is the overall accuracy and class-wise accuracy:

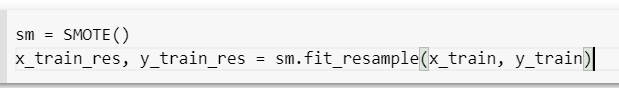




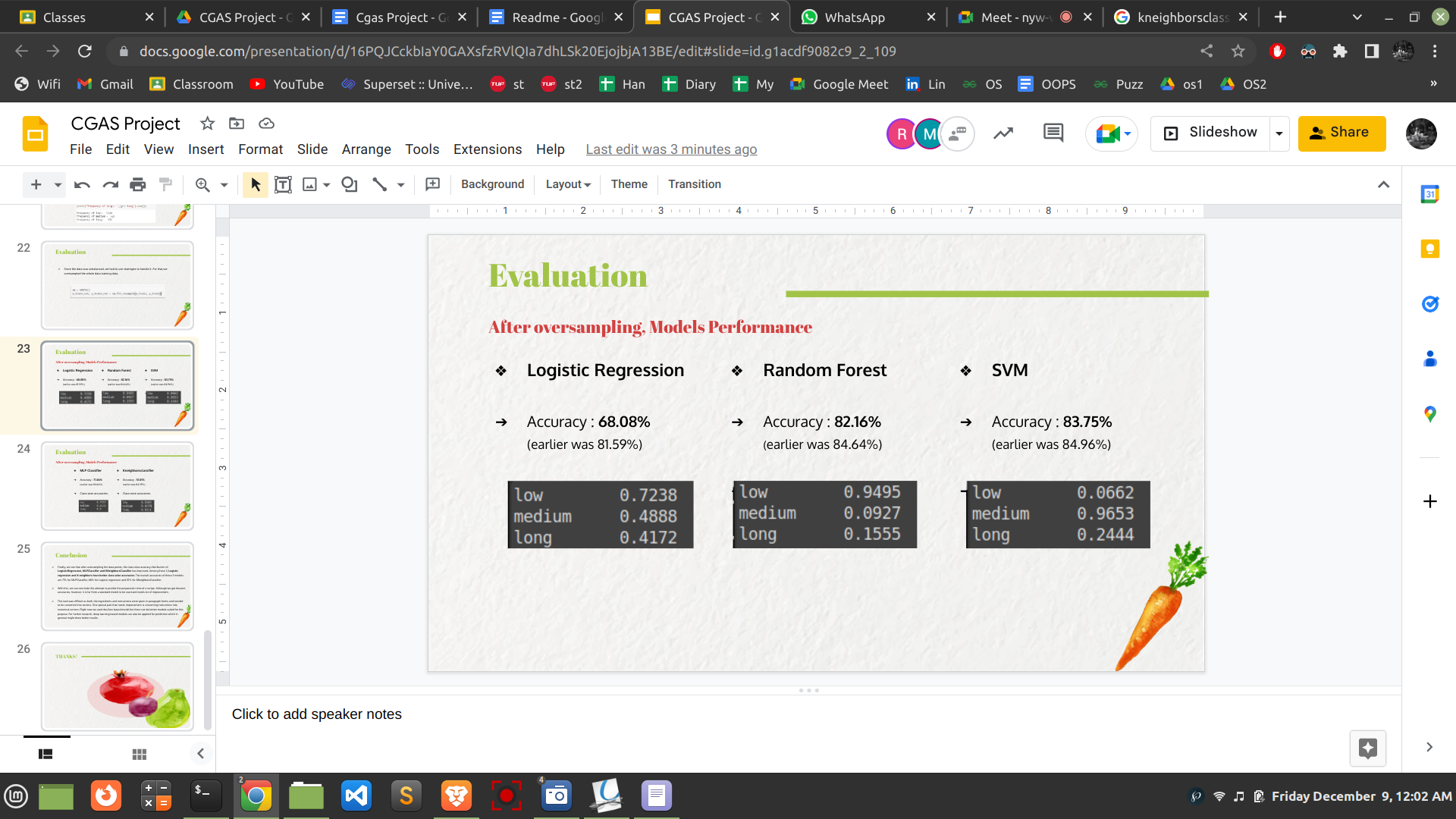
From above results, we see that although the overall accuracy is high, however the class-wise accuracies are poor for 'medium' and 'long' classes. This happened because the 3 classes are imbalanced.

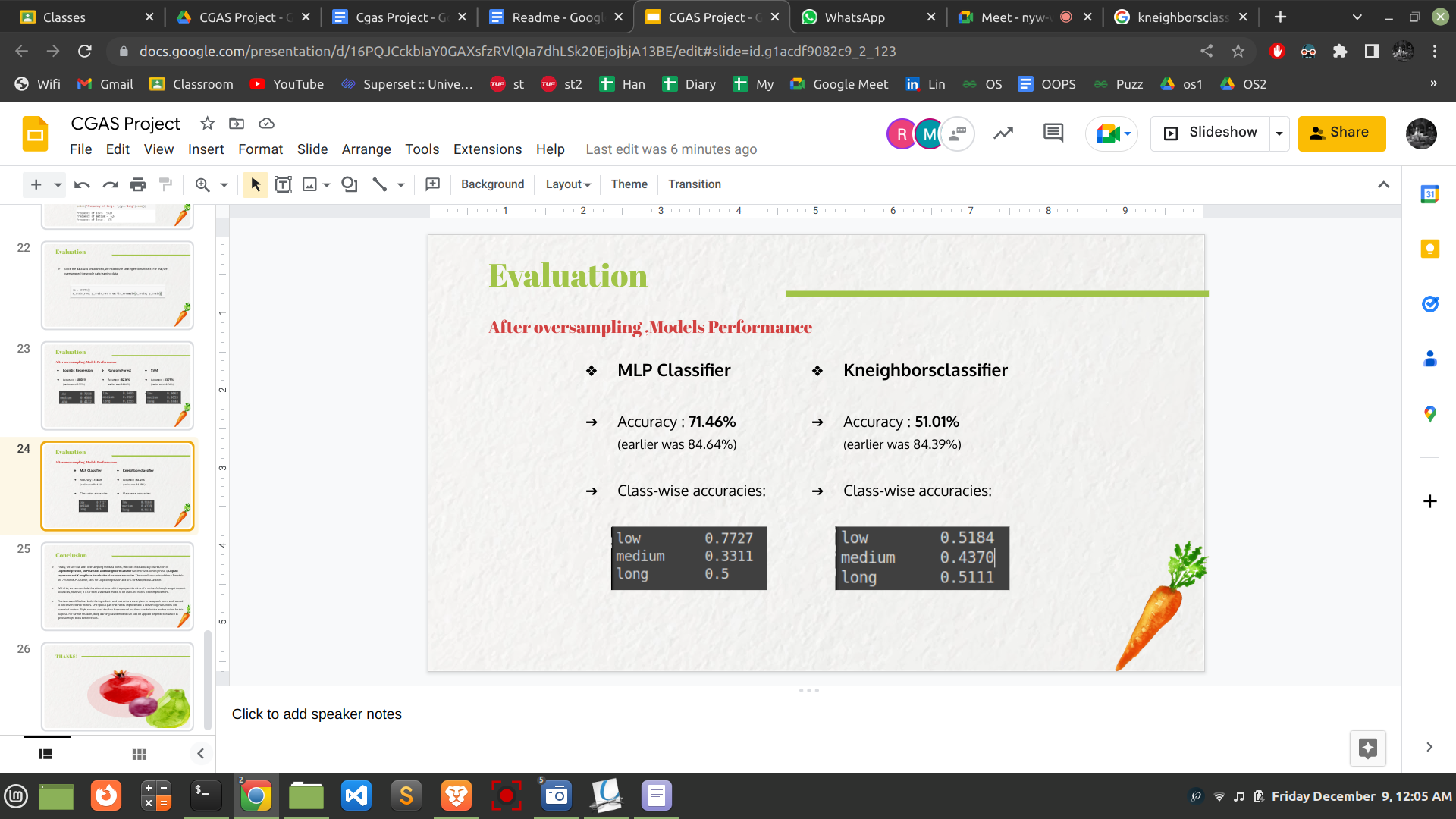
We can check that by seeing the frequencies of the 3 classes:



Since the data was unbalanced, we had to use strategies to handle it. For that,we oversampled the whole data training data. 

After oversampling,the accuracies changed as shown below:





**Conclusion**

* Finally, we see that after oversampling the data points, the class-wise accuracy distribution of **LogisticRegression, MLPClassifier and KNeighborsClassifier** has improved. Among these 3, **Logistic regression and K-neighbors have better class-wise accuracies.** The overall accuracies of these 3 models are 71.46% for MLPClassifier, 68.08% for Logistic regression and 51.01% for KNeighborsClassifier.
* With this, we can conclude this attempt to predict the preparation time of a recipe. Although we got decent accuracy, it is far from a standard model to be used and needs a lot of improvement.
* This task was difficult as both the ingredients and instructions were given in paragraph forms and needed to be converted into vectors. One special part that needs improvement is converting instructions into numerical vectors. Right now we used doc2vec based model but there can be better models suited for this purpose. For further research, deep learning based models can also be applied for prediction which in general might show better results.