Food Classification

Machine Learning - Mini Project

Team Members:

Vishnu K Krishnan	-	18 5001 200
Shashanka Venkatesh	-	18 5001 145
Suraj Jain	-	18 5001 177
Vishakan Subramanian	-	18 5001 196

About the Dataset

Dataset

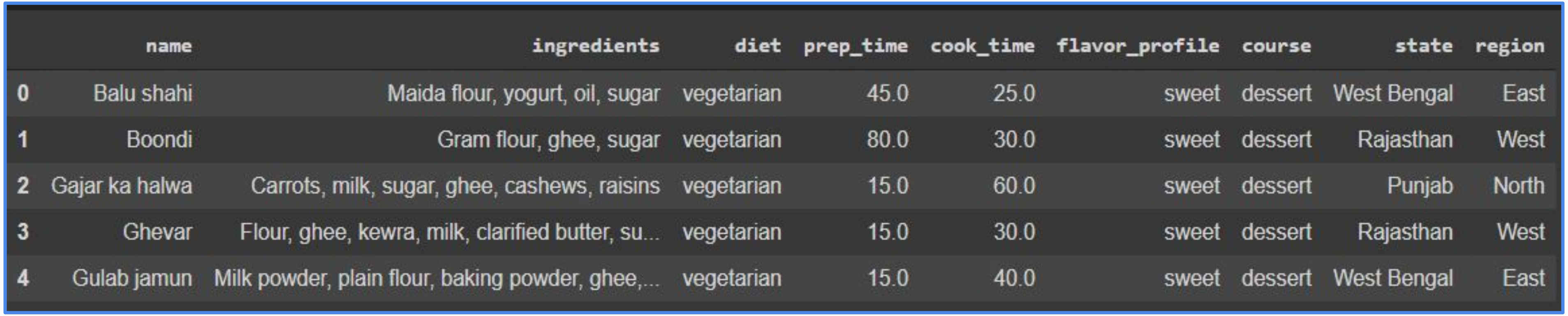
The dataset was obtained from from Kaggle: Indian Food 101. It has **255** different Indian dishes, described by **9 features** namely:

1. Name of the Dish
2. Ingredients Used in the dish
3. Diet (Veg/Non-Veg)
4. Preparation Time
5. Cooking Time
6. Flavor Profile
7. Course (Starter/Main/Dessert)
8. State
9. Region.

Since our task is food classification, our target variable is **Diet**, and we would like to classify whether a dish is **vegetarian or non-vegetarian,** given the other features.

Dataset

Sample training examples from the dataset



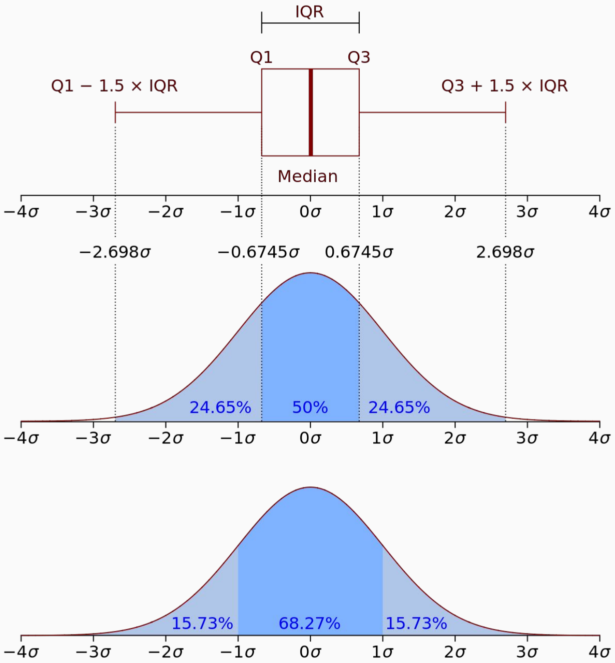
	name	ingredients	diet	prep_time	cook_time	flavor_profile	course	state	region
0	Balu shahi	Maida flour, yogurt, oil, sugar	vegetarian	45.0	25.0	sweet	dessert	West Bengal	East
1	Boondi	Gram flour, ghee, sugar	vegetarian	80.0	30.0	sweet	dessert	Rajasthan	West
2	Gajar ka halwa	Carrots, milk, sugar, ghee, cashews, raisins	vegetarian	15.0	60.0	sweet	dessert	Punjab	North
3	Ghevar	Flour, ghee, kewra, milk, clarified butter, su...	vegetarian	15.0	30.0	sweet	dessert	Rajasthan	West
4	Gulab jamun	Milk powder, plain flour, baking powder, ghee,...	vegetarian	15.0	40.0	sweet	dessert	West Bengal	East

Data Preprocessing

Outlier Handling

In order to remove outliers from the dataset, we are proposing to use **Interquartile range (IQR).**

It is a simple statistic value that works on the assumption that the data is normally distributed. The theory is that **~97%** of all data lie within **2 Standard Deviations from the mean.** Therefore, removing data points outside that region will get rid of the outliers in the dataset.

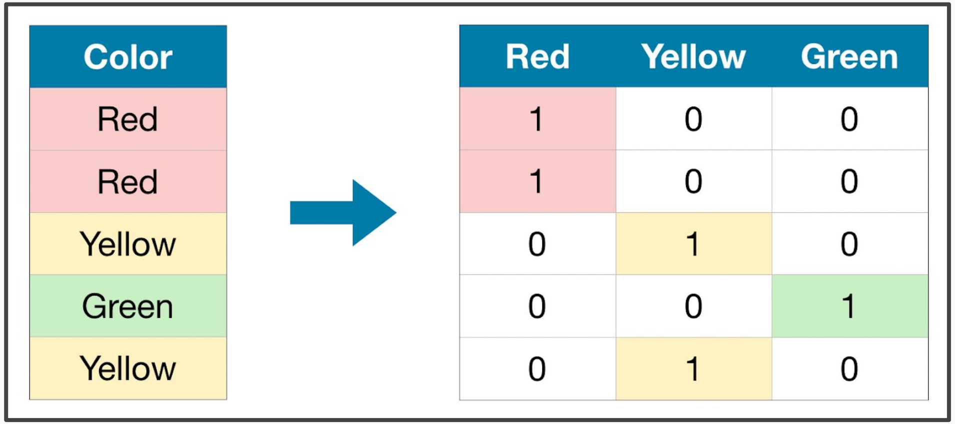


One-Hot Encoding

We have proposed to use **One-Hot Encoding** on the categorical features of the dataset. This allows the dataset to be fed into a Neural Network.

On One-Hot Encoding is performed on the **flavour profile, course, region & state** features.This expands the input vector size to **40.**

On including the **ingredients** feature, the input vector size blows up to **400**.



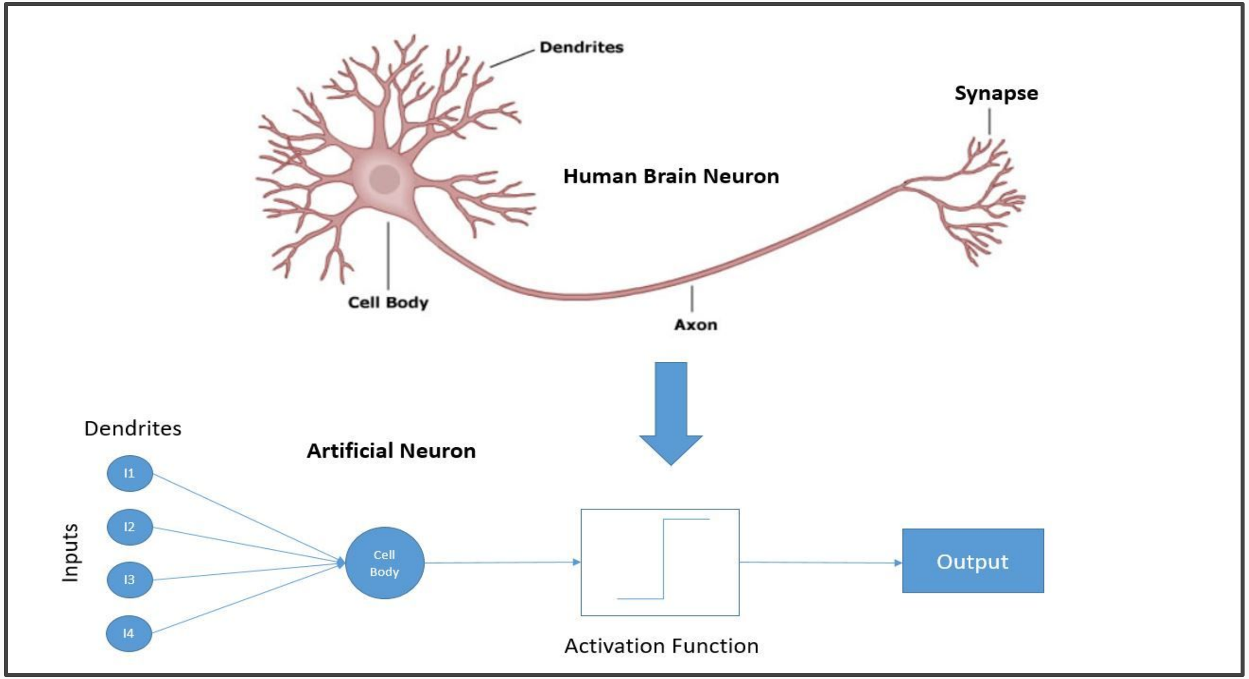
Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0

Sample One-Hot Encoding of Categorical Data

Our Proposed Architecture

Artificial Neural Network

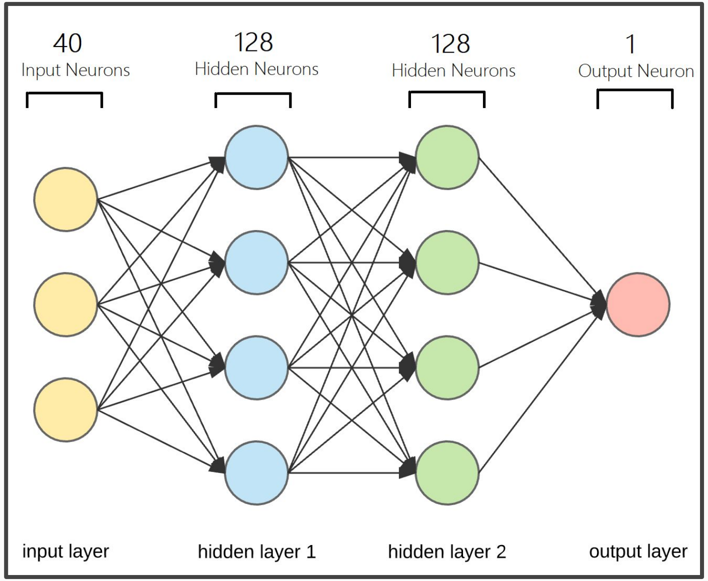


Artificial Neural Network

Model Proposed

We propose to use an Artificial Neural Network with 2 hidden layers to perform the classification that we desire. The inputs are the pre-processed data values, and the output is a single value between 0 and 1.

The proposed architecture is a **[40 x 128 x 128 x 1] Neural Network**, with a **ReLu** activation function for both the hidden layers and a **Sigmoid** activation function for the output layer.



Loss Function

For our loss function, we believe that the **Binary Cross-Entropy** loss is the most suitable for our task, since the output feature is binary (vegetarian/non-vegetarian)

It is also known as the **log-loss**, and is calculated in the following manner:

$$\boxed{\text{Loss} = -\frac{1}{\substack{\text{output}\\\text{size}}} \sum_{i=1}^{\substack{\text{output}\\\text{size}}} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log\,(1 - \hat{y}_i)}$$

Performance Metrics

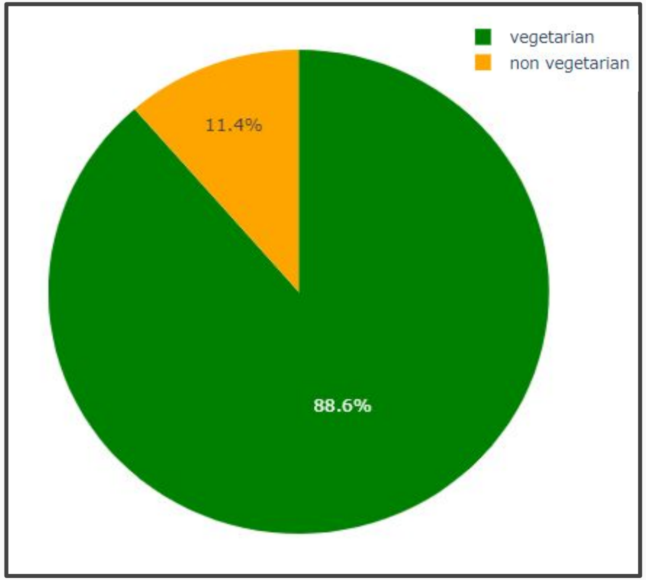
Metrics To Be Used

We plan to use the following metrics:

- ➔ Validation Loss & Validation Accuracy
- ➔ Training & Testing Accuracy
- ➔ AUC (Area Under the Receiver-Operator Curve)
- ➔ Classification Report (with F1 score, Precision, Recall etc.)
- ➔ Confusion Matrix, Matthew's Correlation Coefficient, Balanced Accuracy Score

Distribution of the Data

The dataset is heavily imbalanced, with only **11.4%** of the data being non-vegetarian dishes, while the remaining **88.6%** are vegetarian dishes. Thus, we need to verify our model's accuracy with metrics like **Matthew's Correlation Coefficient** and **Balanced Accuracy Score**, to make sure that the model is performing upto mark, and is not biased.



Going Further

Random Forest Classifier

Random Forest is a robust machine learning model that has gained popularity in the recent years.

We would also like to use the Random Forest Classifier to perform the required machine learning task, and perform a comparison between the ANN and the RF model.

Random Forest is a meta classifier that fits a number of decision tree classifiers on various sub-samples of the data and uses the principle of averaging and **bagging** (bootstrap aggregating) to improve the overall accuracy of the model.

Thank You!