DIETARY ASSESSMENT AND INDIAN CUISINE ANALYSIS USING KNN AND EDA

[1]Pearl Ahuja, [2]Diksha Solanki, [3]Amita Goel, [4]Nidhi Sengar, [5]Vasudha Bahl

[1]Student, [2]Student, [3]Teacher, [4]Teacher, [5]Teacher
[1]Information Technology,
[1]Maharaja Agrasen Institute of Technology, Delhi, India

Abstract: Indian Cuisine consists of a variety of regional and traditional cuisines native to the Indian Subcontinent. Given the diversity in soil, climate, culture, ethnic groups, and occupations, these cuisines vary substantially and use locally available spices, herbs, vegetables, and fruits. Indian food is also heavily influenced by religion, cultural choices and traditions.

The cuisine of India is one of the world's most diverse cuisines, characterized by its sophisticated and subtle use of the many spices, vegetables, grains and fruits grown across India. The cuisine of each geographical region includes a wide assortment of dishes and cooking techniques reflecting the varied demographics of the ethnically diverse Indian subcontinent. The project is based on exploratory data analysis on data set 'Indian Food 101'. The data set contains the information about famous Indian dishes.

We plot some graphs to know the trends of food in different region and correlation between cooking time and food item, number of food items for each state veg-nonveg and recommendation of similar dishes using K-nearest neighbor.

Index Terms - Dietary Assessment · K-Nearest Neighbors · Food analysis · Food Classification · Cuisine Analysis

I. INTRODUCTION

India is unique in its own way. It is not confined to one culture or one language, but has several cultures flowing through its vast lands and many languages are spoken. It is multi-cultural and multi-lingual unit, encompassing people from various social and ethnic backgrounds. Citizens of India dress up differently, have different cuisine and their social and religious pursuits vary. Indian cuisine varies from region to region. Traditionally, some states in India have their own unique dishes, which they often prepare during religious and social gatherings.

The many similarities between the culinary regions of India are highlighted with an exquisite use of spices and flavorings. These range from cardamom, cumin, cloves, fennel seeds and garlic to ginger, chilies, fenugreek, saffron and turmeric. Spice mixtures or masalas are a crucial element of Indian cuisine. Whether fresh or dried, masalas make use of local ingredients and are prepared daily along with grains, pulses and vegetables.

While meat, poultry, fish and seafood dishes are offered throughout India, most vegetarian specialties are found in the central and southern regions. For an added dimension of flavor, a variety of fruits are served fresh or pickled, such as chutney and relish. Dairy products, such as clarified butter (ghee), cheese (paneer), yogurt, milk and buttermilk (moru)are used as ingredients and condiments. Raita is popular chilled yogurt condiment garnished with chopped fruit or vegetables and spices.

Dal preparations (dried legumes and pulses) are at the center of Indian meals. When combined with grains, they provide an inexpensive source of essential protein. Dals are prepared whole and pureed and are generally served with vegetables and meat, where accepted. In the northern regions, thick and hearty stew-like dals are eaten with bread, while the thinner preparations of the south are best suited for rice. Channa dal or gram lentils are the most widely grown dal in India. Used both as protein and starch, garbanzos and lentils supply the base for breads, crepes and thickeners for curries. Other types of dals commonly eaten include peas, kidney beans, mung beans and split peas.

Served at nearly every meal, vegetables are one of India's most significant ingredients. India's perfected vegetable cookery offers rich and flavorful dishes, ranging from appetizers and side dishes to entrees and fried, roasted, braised, sautéed, pureed and stuffed. Dairy products, fruits, nuts, spices and seasonings are used to embellish greens (palak),eggplant, gourds, roots and squash while caramelized onions and tomatoes provide the foundation for many sauces and stews. Cauliflower and potatoes (aloo gobhi), peas and potatoes(aloo matter), peas and cheese (matter paneer), and spinach and cheese (saag paneer) are popular vegetable combinations.

There are several types of rice grown and eaten in India, from long-grains and medium-grain to glutinous and wild. While basmati is generally reserved for special occasional, plain boiled rice is served with everyday meals, especially in the southern

areas. One- pot rice dishes, such as biryani, a combination of basmati rice, meat or seafood, vegetables and expensive spices, nuts, fruits and meat, seafood or yogurt are oftentimes prepared for celebrations and religious festivals.

In this paper, an exertion has been made to analyze the various food items for assessment of most favored cuisine in every region of India along with recommendation of various dishes preferable to that region using KNN and EDA. The standard Indian0020Food-101 dataset has been chosen as the working information base for this project.

II. LITERATURE REVIEW

Food and taste are resolute by its culture, anatomy, and genetics. Almost every eatable which are consumed by humans are associated with some of its tradition so as to explore taste and uniqueness. Some people pays top dollar for escargot in fine restaurants while others stomp on the same snail when they find it in the garden. One person's haute cuisine is another person's pest. A destination's local cuisine is "deeply rooted in a particular place, space, and time, its culinary traditions reveal the character of the society and mentality of it members" (Bessiere, 1998, p. 28).

As the world cuisine is becoming increasingly popular in India we also need our local food to be pushing traditional boundaries with a vision of making Indian Cuisine a world Healthiest Cuisine. (Prem Ram and Sonia Sharma (2015).
Cuisine is inextricably linked to the destination in terms of its cultural heritage, political, social, and economic identity. As such, cuisine plays an important role in establishing the destination's overall tourism image (Crofts, 2010; Everett & Aitchison, 2008; Kivela & Crotts, 2005; Lockie, 2001).

Indian cuisine has gained a primary place, especially in the Western world as a result of globalization and other factors such as immigration, availability of recipes on the web and increased tourism activities. From „chicken
tikka masala" becoming the national dish of Britain to many Indian recipes appearing on various international flights, Indian food items have secured their place on the new global menu. Indian cuisine has evolved over the years and it has a strong connection to its culture, history, and geography.

The dietary patterns have also evolved based on various religious practices. (Mangalassary, 2015). Cultural shifts in culinary behavior can be caused by such changes as "male out-migration, inter-class rivalry and imitation, changing caste relations (in India), and market conditions" (Mintz & Du Bois, 2002, p. 104).
More and more Indians today are global citizen who embrace global trends. There is a major culinary revolution going on in the capital with provides ample Cuisine for thought. Japanese Cuisine has taken a lead and Sushi Counters are found in each and every market.
Continental, Thai Spanish, and Italian cuisine is what the younger generation demands says Hussain, S.

III. DATASET

For our purposes, we picked from challenging Indian Food-101 dataset, because we considered them to have the most distinct and representative properties. This dataset consists of 101 categories and each category has 255 rows, in total 9 columns.
Most of the data is about popular Indian food. Following are the attributes we selected for the purpose of the analysis
:
   i.   Name : name of the dish
   ii.   Ingredients : main ingredients used
   iii. Diet : type of diet - either vegetarian or non-vegetarian
   iv.   Prep time : preparation time
   v.   Cook time : cooking time
   vi.   Flavor profile : flavor profile includes whether the dish is spicy, sweet, bitter, etc.
   vii. Course : course of meal - starter, main course, dessert, etc.
   viii. State : state where the dish is famous or is originated
   ix.   Region : region where the state belongs

Indian cuisine consists of a variety of regional and traditional cuisines native to the Indian subcontinent. Given the diversity in soil, climate, culture, ethnic groups, and occupations. These are the top 5 rows representing the features of the defined attributes.

	name	ingredients	diet	prep_time	cook_time	flavor_profile	course	state	region
0	Balu shahi	Maida flour, yogurt, oil, sugar	vegetarian	45	25	sweet	dessert	West Bengal	East
1	Boondi	Gram flour, ghee, sugar	vegetarian	80	30	sweet	dessert	Rajasthan	West
2	Gajar ka halwa	Carrots, milk, sugar, ghee, cashews, raisins	vegetarian	15	60	sweet	dessert	Punjab	North
3	Ghevar	Flour, ghee, kewra, milk, clarified butter, su...	vegetarian	15	30	sweet	dessert	Rajasthan	West
4	Gulab jamun	Milk powder, plain flour, baking powder, ghee,...	vegetarian	15	40	sweet	dessert	West Bengal	East

Figure 1: Dataset

IV. METHODS

Before describing the architecture and the different components of the proposed system, we provide a brief introduction to the KNN and EDA.

K-Nearest Neighbors Algorithm

In statistics, the k-nearest neighbors algorithm (k-NN) is a non-parametric classification method first developed by Evelyn Fix and Joseph Hodges in 1951, and later expanded by Thomas Cover. It is used for classification and regression. In both cases, the input consists of the k closest training examples in data set. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors.

k-NN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically.

Both for classification and regression, a useful technique can be to assign weights to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor.

The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

Exploratory Data Analysis

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.

V. METHODOLOGY

The major topics to be covered are below:

  i.   Handle Missing value
  ii.   Removing duplicates
  iii. Outlier Treatment
  iv.   Normalizing and Scaling(Numerical Variables)
  v.   Encoding Categorical variables(Dummy Variables)
  vi.   Bivariate Analysis

We can see that we have various missing values in the respective columns. There are various ways of treating your missing values in the data set. And which technique to use when is actually dependent on the type of data you are dealing with.

Drop the missing values: In this case, we drop the missing values from those variables. In case there are very few missing values you can drop those values.
Impute with mean value: For the numerical column, you can replace the missing values with mean values. Before replacing with mean value, it is advisable to check that the variable shouldn't have extreme values .i.e. outliers.
Impute with median value: For the numerical column, you can also replace the missing values with median values. In case you have extreme values such as outliers it is advisable to use the median approach.
Impute with mode value: For the categorical column, you can replace the missing values with mode values i.e. the frequent ones.

Sklearn. neighbors provides functionality for unsupervised and supervised neighbors-based learning methods. Unsupervised nearest neighbors is the foundation of many other learning methods, notably manifold learning and spectral clustering. Supervised neighbors-based learning comes in two flavors: classification for data with discrete labels, and regression for data with continuous labels.

The principle behind nearest neighbor methods is to find a predefined number of training samples closest in distance to the new point, and predict the label from these. The number of samples can be a user-defined constant (k-nearest neighbor learning), or vary based on the local density of points (radius-based neighbor learning). The distance can, in general, be any metric measure: standard Euclidean distance is the most common choice. Neighbors-based methods are known as non-generalizing machine learning methods, since they simply "remember" all of its training data (possibly transformed into a fast indexing structure such as a Ball Tree or KD Tree).

Despite its simplicity, nearest neighbors has been successful in a large number of classification and regression problems, including handwritten digits and satellite image scenes. Being a non-parametric method, it is often successful in classification situations where the decision boundary is very irregular.

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of Matplotlib.
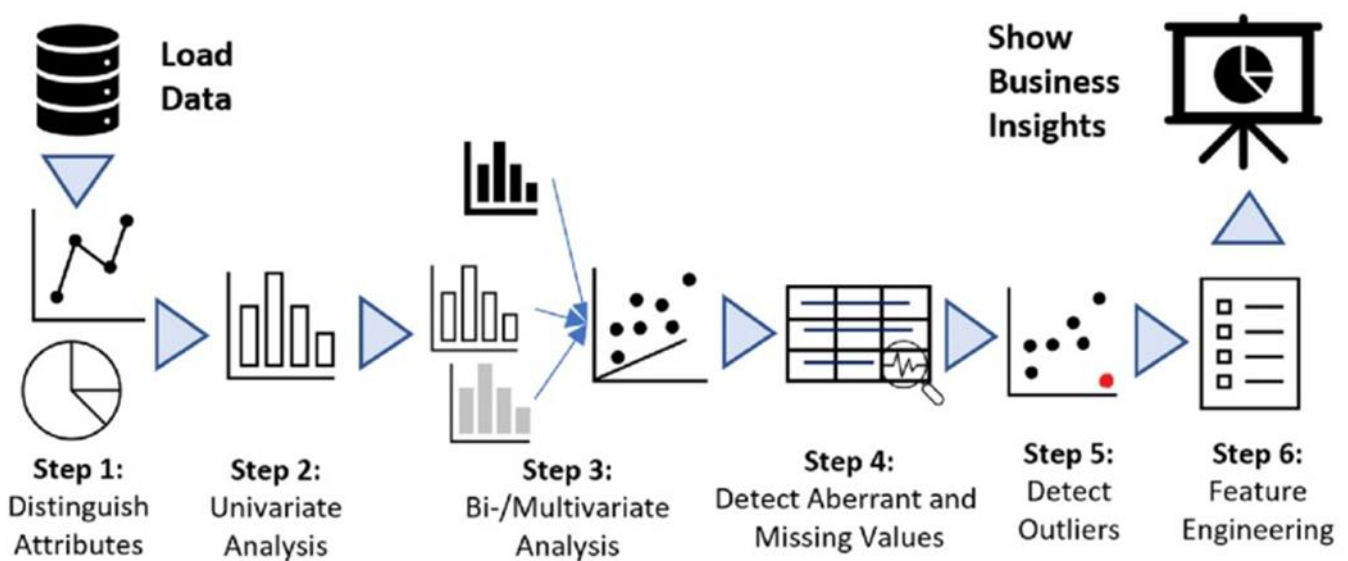

Figure 2: Exploratory Data Analysis

VI. EXPERIMENTAL RESULTS

Recommendation for chicken tikka masala are following:

i. Shahi paneer with distance 0.19821627426272703
ii. Chana masala with distance 0.19821627426272703
iii. Bhindi masala with distance 0.19821627426272703
iv. Pindi chana with distance 0.2284832501895404
v. Kadai paneer with distance 0.2284832501895404
vi. Palak paneer with distance 0.2284832501895404
vii. Paneer butter masala with distance 0.2284832501895404
viii. Dal tadka with distance 0.2857142857142859
ix. Makki di roti sarson da saag with distance 0.2857142857142859
x. Chole bhature with distance 0.2857142857142859

Inferences:
➢ Most of the ingredients are different in vegetarian and non-vegetarian dishes
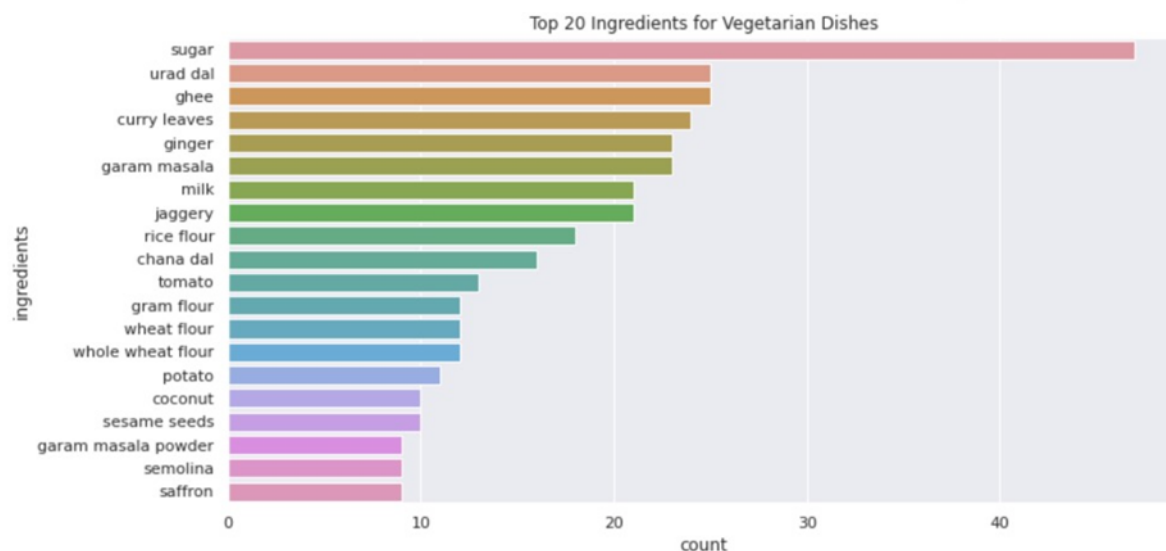➢ Veg and non-veg dishes take a similar amount of time
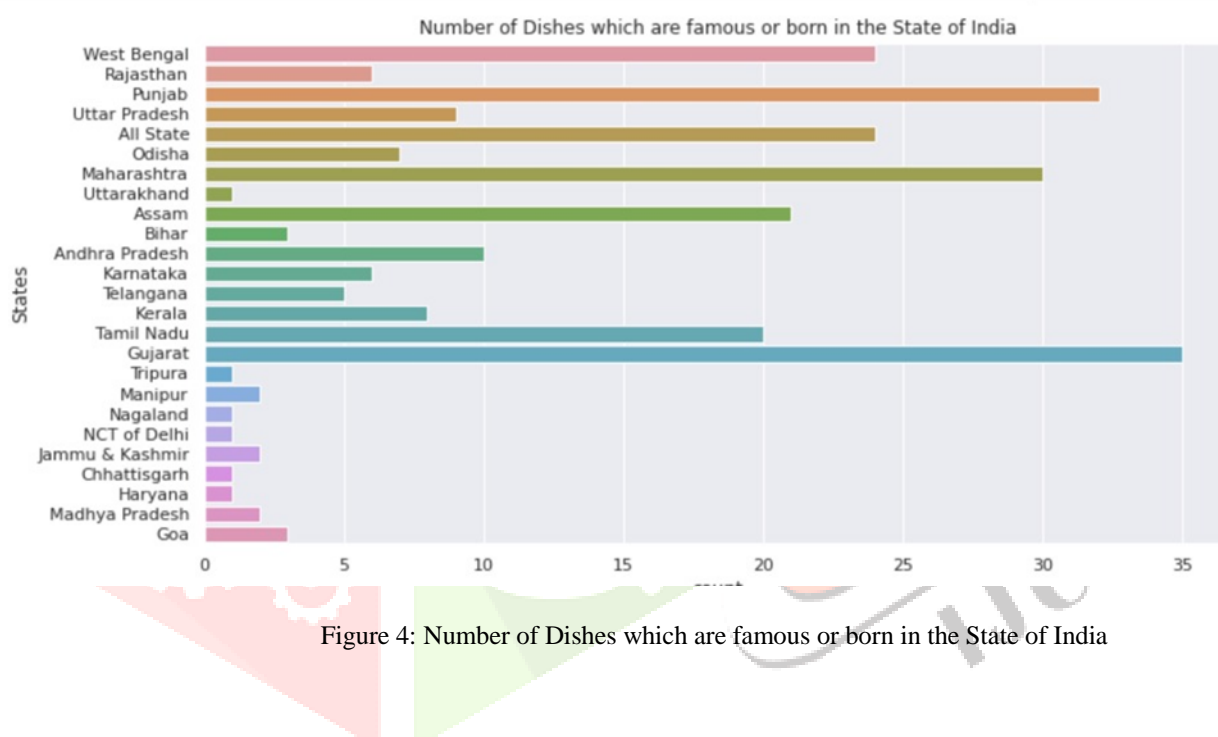
Figure 3: Ingredients for Vegetarian Dishes

Figure 4: Number of Dishes which are famous or born in the State of India

VII. CONCLUSION

The Following are the Specialty Foods from different States:

i. Misti doi - West Bengal
ii. Ras malai - West Bengal
iii. Obbattu holige - Karnataka
iv. Poornalu - Andhra Pradesh
v. Chak Hao Kheer - Manipur
vi. Chicken Tikka - Punjab
vii. Pindi chana - Punjab
viii. Tandoori Chicken - Punjab
ix. Tandoori Fish Tikka - Punjab
x. Attu - Andhra Pradesh
xi. Dosa - Tamil Nadu
xii. Idiappam - Tamil Nadu
xiii. Idli - Andhra Pradesh
xiv. Masala Dosa - Andhra Pradesh
xv. Pesarattu - Andhra Pradesh
xvi. Puttu - Kerala
xvii. Sandige - Karnataka
xviii. Sevai - Andhra Pradesh

xix.   Malapua - Biha
xx.   Malapua - Bihar
xxi.   Shrikhand - Maharashtra
xxii. Biryani - Hyderabad/Telangana
xxiii. Pindi chana – Punjab

VIII. REFERENCES

1. Chen, M.; Dhingra, K.; Wu, W.; Yang, L.; Sukthankar, R.; Yang, J. PFID: Pittsburgh Fast Food Image Dataset. In Proceedings of the ICIP 2009, Cairo, Egypt, 7–10 November 2009;

2. Chen, M. Y., Yang, Y. H., Ho, C. J., Wang, S. H., Liu, S. M., Chang, E., & Ouhyoung, M. (2012, November). Automatic chinese food identification and quantity estimation. In SIGGRAPH Asia 2012 Technical Briefs (p. 29). ACM.

3. Lowe, D.G. Object Recognition from Local Scale Invariant Features. In Proceedings of the ICCV'99, Corfu, Greece, 20–21 September 1999; pp. 1150–1157.

4. Yanai, K., & Kawano, Y. (2015, June). Food image recognition using deep convolutional network with pre-training and fine-tuning. In Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on (pp. 1-6). IEEE.

5. Kagaya, H.; Aizawa, K.; Ogawa, M. Food Detection and Recognition using Convolutional Neural Network. In Proceedings of the MM'14, Orlando, FL, USA, 3–7 November 2014; pp. 1055– 1088.

6. Probst, Y., Nguyen, D. T., Tran, M. K., & Li, W. (2015). Dietary assessment on a mobile phone using image processing and pattern recognition techniques: Algorithm design and system prototyping. Nutrients, 7(8), 6128-6138.

.