

Data Collection and Preprocessing Phase

| | |
|---------------|----------------------------------|
| Date | 26 May 2025 |
| Name | Shambhuraj Jaysing Patil |
| Project Title | Restaurant Recommendation System |
| Maximum Marks | 6 Marks |

Data Preprocessing

The images will be preprocessed by resizing, normalizing, augmenting, denoising, adjusting contrast, detecting edges, converting color space, cropping, batch normalizing, and whitening data. These steps will enhance data quality, promote model generalization, and improve convergence during neural network training, ensuring robust and efficient performance across various computer vision tasks.

| Section | Description |
|-------------------|--|
| Data Overview | The dataset contains restaurant information from Zomato, including name, reviews, ratings, cuisines, cost, and more. The data is cleaned, deduplicated, and preprocessed for building a content-based recommendation system. |
| Resizing | <i>Not applicable for text data.</i> |
| Normalization | Ratings are normalized to a 1-5 scale using MinMaxScaler. Text is lowercased and punctuation is removed. |
| Data Augmentation | Not applicable for text data. |
| Denoising | Text is cleaned by removing newline characters and punctuation. |
| Edge Detection | Not applicable for text data. |

| | |
|--|--|
| Color Space Conversion | Not applicable for text data. |
| Image Cropping | Not applicable for text data. |
| Batch Normalization | Not applicable for text data. |
| Data Preprocessing Code Screenshots | |
| Loading Data | <pre> # Mounting Google Drive #from google.colab import drive #drive.mount('/content/drive') import csv # Specifying the path to the dataset file file_path = '/content/zomato.csv' # Reading the dataset into a Pandas DataFrame #df = pd.read_csv(file_path,encoding = 'ISO-8859-1', low_memory = False) df = pd.read_csv(file_path, encoding='ISO-8859-1', on_bad_lines='skip', engine='python') # Displaying the first few rows of the dataset to ensure it's loaded correctly df.head()</pre> |
| Resizing | <i>Not applicable</i> |
| Normalization | <pre> # Computing Mean Rating restaurants = list(df['name'].unique()) df['Mean Rating'] = 0 for i in range(len(restaurants)): df['Mean Rating'][df['name'] == restaurants[i]] = df['rate'][df['name'] == restaurants[i]].mean() #Scaling the mean rating values from sklearn.preprocessing import MinMaxScaler scaler = MinMaxScaler (feature_range = (1,5)) df[['Mean Rating']] = scaler.fit_transform(df[['Mean Rating']]).round(2)</pre> |
| Data Augmentation | <i>Not applicable</i> |
| Denoising | <pre> ## Lower Casing df["reviews_list"] = df["reviews_list"].str.lower() ## Removal of Punctuations import string PUNCT_TO_REMOVE = string.punctuation def remove_punctuation(text): """custom function to remove the punctuation""" return text.translate(str.maketrans('', '', PUNCT_TO_REMOVE)) df["reviews_list"] = df["reviews_list"].apply(lambda text: remove_punctuation (text))</pre> |
| Edge Detection | <i>Not applicable</i> |

| | |
|------------------------|-----------------------|
| Color Space Conversion | <i>Not applicable</i> |
| Image Cropping | <i>Not applicable</i> |
| Batch Normalization | <i>Not applicable</i> |