Predictive Modeling Report

Understanding Online Food Ordering Behavior

RADHIKA GUPTA
05001192022
B.Tech in AIML(4th Sem)
ML PROJECT

PROBLEM STATEMENT

This analysis aims to understand the drivers of online food ordering and create predictive models for customer preferences and feedback. We'll investigate how demographic factors like age, gender, marital status, occupation, and education impact ordering behavior, along with location variables such as latitude, longitude, and pin code. By analyzing feedback, we'll pinpoint areas for service enhancement. Using demographic and location data, we'll develop models to forecast preferences, order status, and feedback. We'll then compare model performance to identify the best approach for predicting online food ordering behavior, offering insights to refine marketing strategies and enhance customer experiences.

key steps

01 DATA COLLECTION

02 DATA PREPARATION

03 SPLITTING THE DATA

04 CHOOSING A MODEL

05 TRAINING THE MODEL

06 EVALUATING THE MODEL

07 DEPLOYMENT

Data Overview

DATASET LINK -

https://www.kaggle.com/datasets/sudarshan24byte/online-food-dataset

```
gdf.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 388 entries, 0 to 387
Data columns (total 22 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Age                                   388 non-null    int64
 1   Monthly Income                        388 non-null    object
 2   Family size                           388 non-null    int64
 3   latitude                              388 non-null    float64
 4   longitude                             388 non-null    float64
 5   Pin code                              388 non-null    int64
 6   Output                                388 non-null    int64
 7   Feedback                              317 non-null    float64
 8   Gender_Female                         388 non-null    uint8
 9   Gender_Male                           388 non-null    uint8
 10  Marital Status_Married                388 non-null    uint8
 11  Marital Status_Prefer not to say      388 non-null    uint8
 12  Marital Status_Single                 388 non-null    uint8
 13  Occupation_Employee                   388 non-null    uint8
 14  Occupation_House wife                 388 non-null    uint8
 15  Occupation_Self Employeed             388 non-null    uint8
 16  Occupation_Student                    388 non-null    uint8
 17  Educational Qualifications_Graduate   388 non-null    uint8
 18  Educational Qualifications_Ph.D       388 non-null    uint8
 19  Educational Qualifications_Post Graduate  388 non-null  uint8
 20  Educational Qualifications_School     388 non-null    uint8
 21  Educational Qualifications_Uneducated  388 non-null    uint8
dtypes: float64(3), int64(4), object(1), uint8(14)
memory usage: 29.7+ KB
```

Data Attributes-

Demographic Information

Age: Age of the customer.

Gender: Gender of the customer.

Marital Status: Marital status of the customer.

Occupation: Occupation of the customer.

Monthly Income: Monthly income of the customer.

Educational Qualifications: Educational qualifications of the customer.

Family Size: Number of individuals in the customer's family.

Location Information

Latitude: Latitude of the customer's location.

Longitude: Longitude of the customer's location.

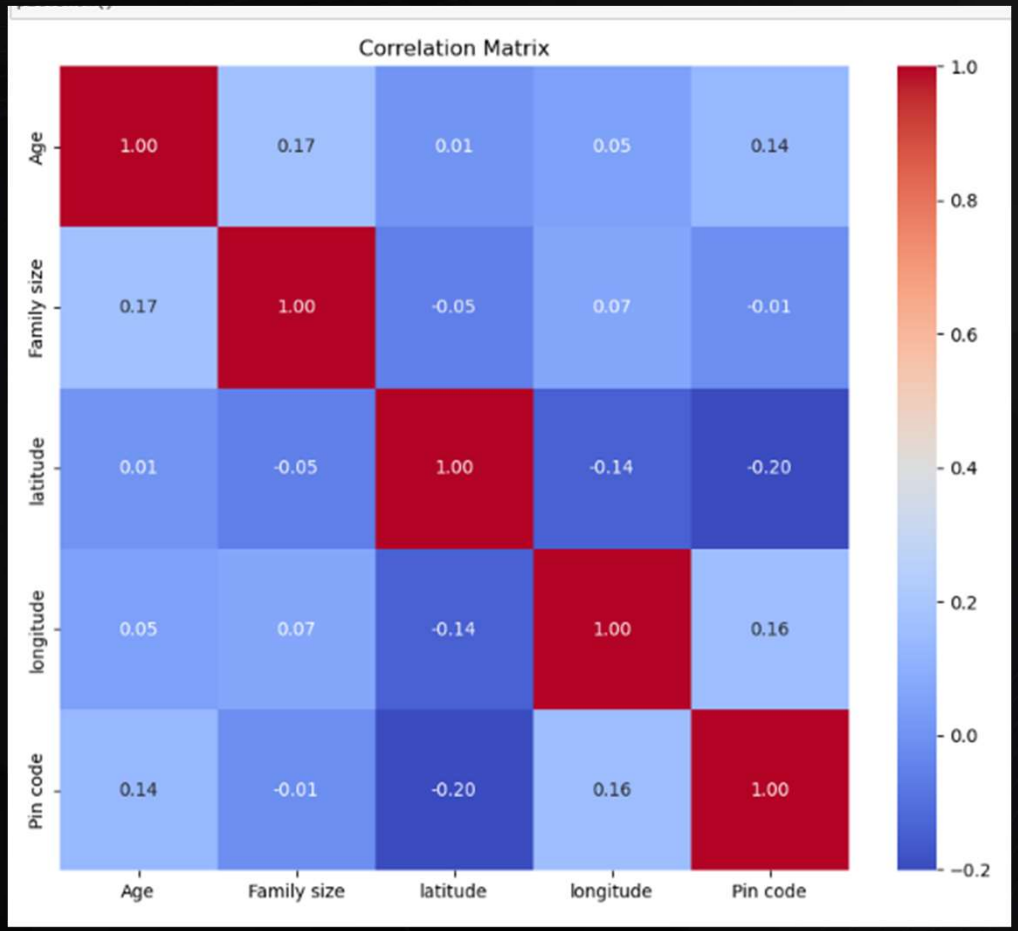
Pin Code: Pin code of the customer's location.

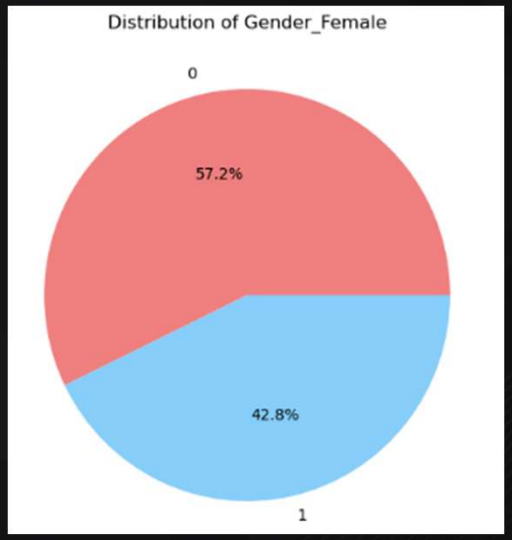
Order Details

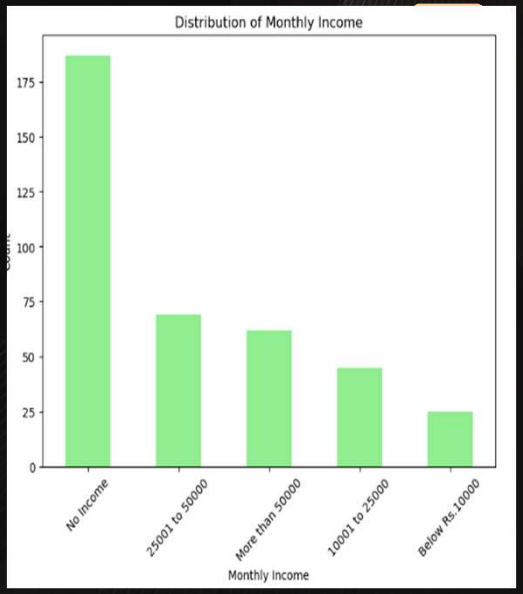
Output: Current status of the order (e.g., pending, confirmed, delivered).

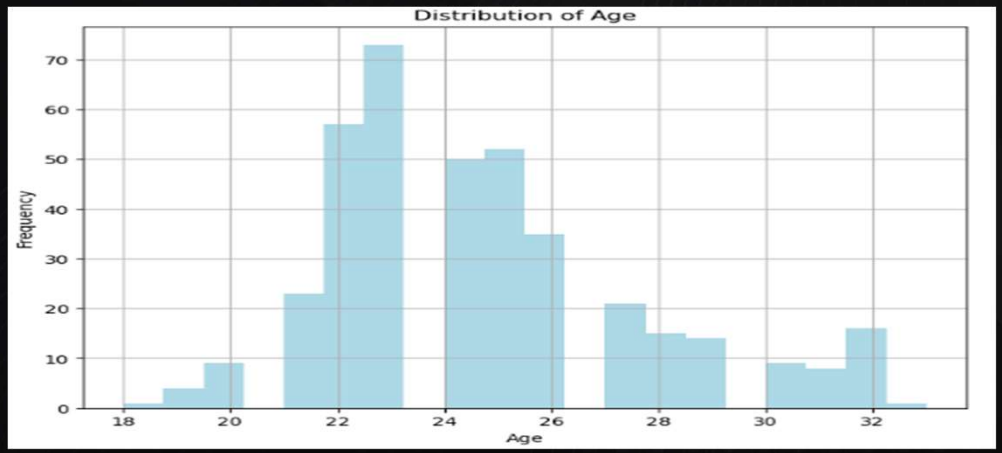
Feedback: Feedback provided by the customer after receiving the order.

Data Visualisation









Data Preparation

```
# Data Preparation
# 1. Handling Missing Values
# Assuming 'Feedback' column contains missing values, we can impute them with the most frequent value
gdf['Feedback'].fillna(gdf['Feedback'].mode()[0], inplace=True)
```

```
#2. Encoding remaining categorical variables
gdf_encoded = pd.get_dummies(gdf, columns=['Feedback'])
# Using one-hot encoding for 'Monthly Income' column
gdf_encoded = pd.get_dummies(gdf, columns=['Monthly Income'])
```

```
# 3. Feature Scaling
# Let's use StandardScaler to scale numerical features
scaler = StandardScaler()
numerical_columns = ['Age', 'Family size', 'latitude', 'longitude', 'Pin code']
gdf_encoded[numerical_columns] = scaler.fit_transform(gdf_encoded[numerical_columns])
```

Data Modelling

1.Linear Regression

```
Mean Squared Error: 0.106326851632879
R-squared: 0.183216457911065
R-squared: 0.3260779839745073
```

Linear regression is a simple and interpretable model that can provide insights into the relationship between the features and the target variable.

2.Random Forest Classifier

```
Random Forest Classifier Evaluation:
Mean Squared Error (MSE): 0.08974358974358974
R-squared (R2) Score: 0.31060606060606033
Root Mean Squared Error (RMSE): 0.29957234475763905
Accuracy Score: 0.9102564102564102
```

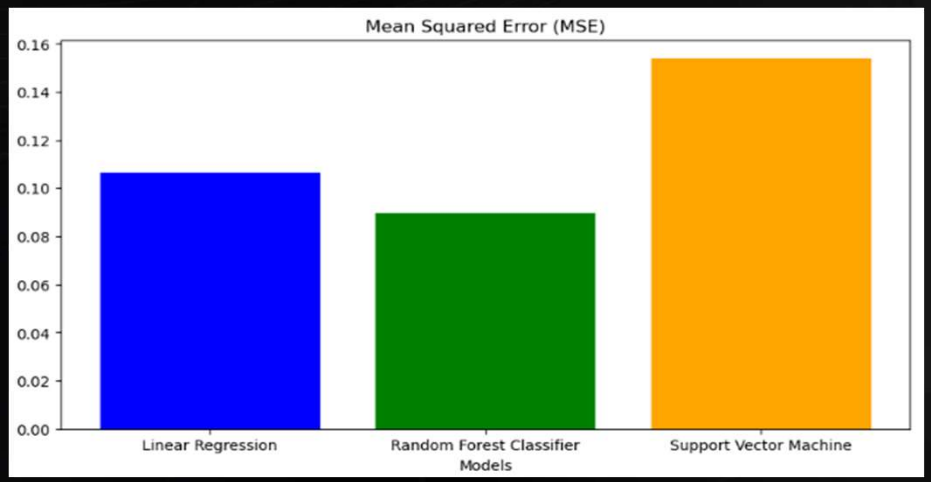
Random forests are versatile and robust ensemble learning methods that can handle non-linear relationships and interactions between features well. They also offer feature importances which can be useful for understanding the data

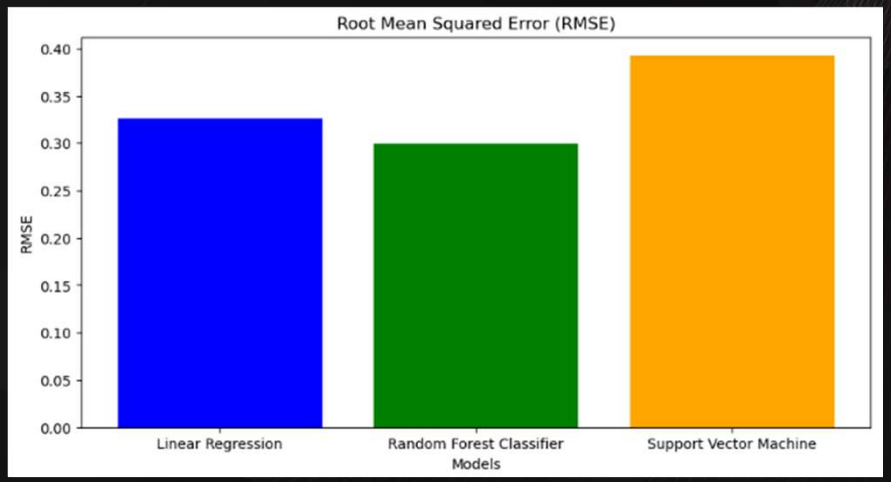
3.Support Vector Machine (SVM)

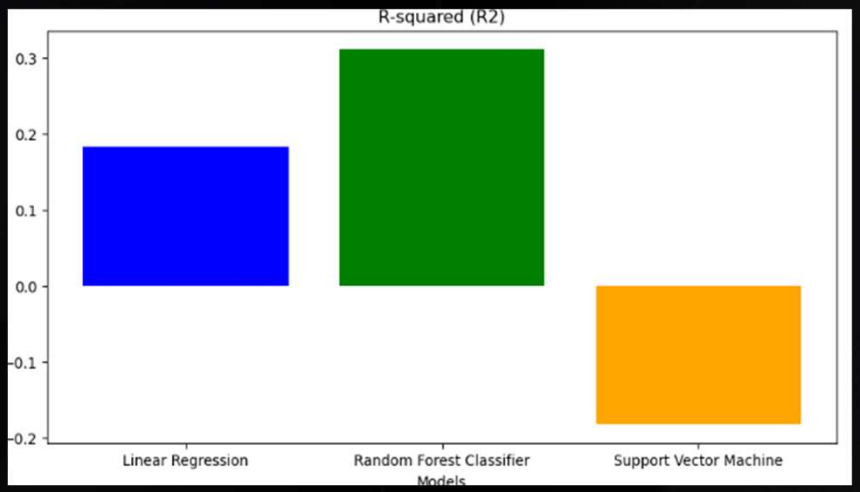
```
Mean Squared Error (MSE) for SVM: 0.1538461538461538
R-squared (R2) for SVM: -0.1818181818181823
Root Mean Squared Error (RMSE) for SVM: 0.3922322702763681
Accuracy Score for SVM: 0.8461538461538461
```

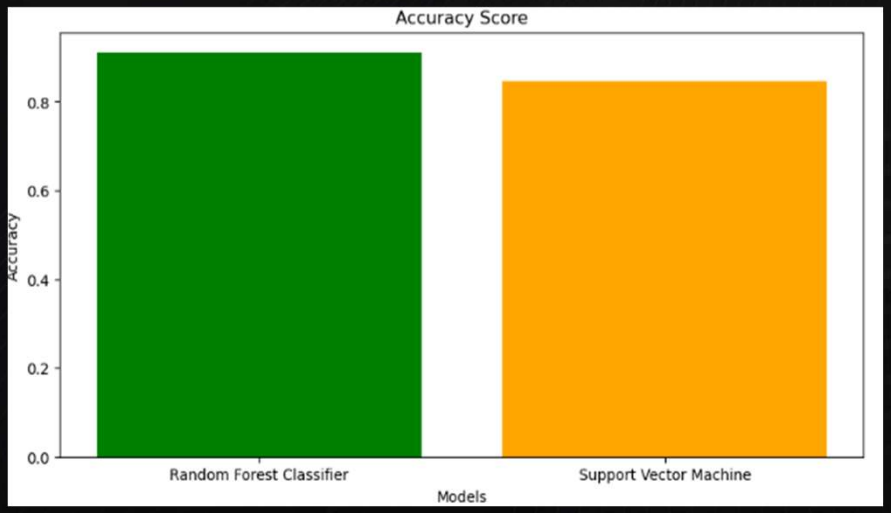
SVMs are powerful models for classification tasks, especially when the data has complex relationships and high dimensionality. Here, we're using a linear kernel for simplicity. SVMs work well when there's a clear margin of separation between classes.

Model Comparison









CONCLUSION

Based on these metrics, the Random Forest Classifier outperforms the other models in terms of both regression metrics (MSE, R-squared, RMSE) and classification accuracy. It has the lowest MSE and RMSE among the three models, indicating better predictive performance in regression tasks. Additionally, it achieves the highest accuracy score, indicating superior classification performance compared to the other models.

NOVELTY IN THE PROJECT

the novelty in this project lies in the comprehensive analysis of online food ordering behavior, incorporating demographic and location factors to develop predictive models. By integrating customer feedback analysis and predictive modeling, the project offers actionable insights for improving service quality and customer satisfaction. The project's holistic approach to understanding customer behavior and its application of machine learning techniques for predictive analytics contribute to its uniqueness.

Reference:-

https://www.researchgate.net/publication/366266285_The_use_of_machine_learning_to_predict_the_main_factors_that_influence_the_continuous_usage_of_mobile_food_delivery_apps