School of Engineering & Technology

Department of Computer Science & Engineering

THE CANTEEN MENU OPTIMIZER

MINI PROJECT REPORT

MACHINE LEARNING CLASSIFICATION CHALLANGE

KARAN PRABHAT
PROGRAM — B.TECH CSE [AI-ML]
ENROLMENT — 2311200010011
REGISTRATION — 230100110730
SECTION — 09 | SEMESTER — 5
SUBJECT — MACHINE LEARNING

SUBMITTED TO — PROF. BIDYUT SAHA

TABLE OF CONTENTS

1. ABSTRACT

This project develops a machine learning model to predict students' **dietary preferences** (Veg, Non-Veg, Vegan, Jain, Eggetarian) with the aim of helping the university canteen manager **reduce food waste and improve menu planning**.

The dataset consisted of **111 student responses** collected via Google Forms. After preprocessing (missing value handling, feature engineering with BMI, and encoding categorical variables), we trained and tuned a **Random Forest Classifier**.

The model achieved strong performance on the **majority class (Non-Veg)** but struggled with minority categories due to severe class imbalance (~85% Non-Veg vs <6% others). **Feature importance analysis** revealed cuisine preference, spice tolerance, and BMI as the most influential features. Despite limitations, the project demonstrates the potential of machine learning for operational decision-making in food services.

2. PROBLEM STATEMENT

Canteens face difficulty in stocking the right proportion of food items, often leading to **food waste** and **shortages**. Our project, the **Canteen Menu Optimizer**, predicts a student's dietary preference based on survey features.

The goal is to:

- ✓ Predict diet type (Veg/Non-Veg/Vegan/etc.)
- ✓ Reduce food waste.
- ✓ Improve student satisfaction.
- ✓ Enable smarter menu planning for the canteen.

3. DATASET DESCRIPTION

- **Size**: 111 rows × 73 columns.
- **Source**: Student responses via Google Forms.
- **Target variable**: dietary_preference.
- **Imbalance**:
  - Non-Veg: ~85%
  - Veg: ~6%
  - Jain, Vegan, Eggetarian: each <5%
- **Features**:
  - Cuisine preferences (categorical, many sparse/missing).
  - Spice tolerance (numeric).
  - Sweet tooth level (numeric).
  - Height & Weight (used to derive BMI).

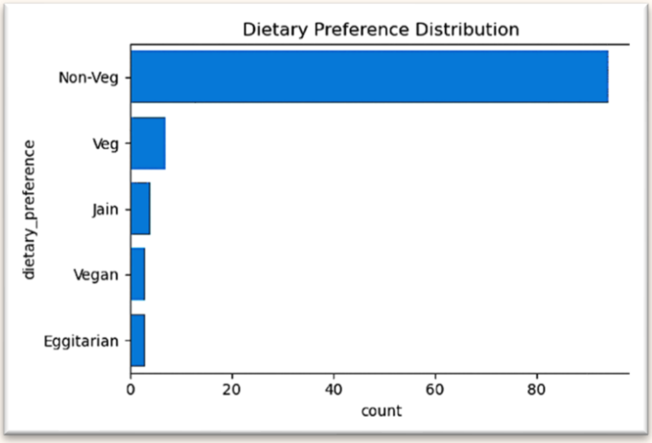

Fig.1 : Distribution of Dietary Preference

4. PREPROCESSING

The dataset required several preprocessing steps:

1) **Cleaning & Transformation**
   - Fixed inconsistencies in age, height, and weight.
   - Dropped irrelevant or redundant columns.
2) **Feature Engineering**
   - Derived **BMI** from height and weight.
   - Created num_cuisines = number of cuisines selected.
   - Added interaction feature (*spice* × *sweet*).
3) **Handling Missing Values**
   - Used SimpleImputer for numeric and categorical features.
4) **Encoding & Scaling**
   - OneHotEncoder for categorical features.
   - StandardScaler for numeric features.

5. EXPLARATORY DATA ANALYSIS (EDA)

EDA revealed key trends:

- Target distribution: strongly dominated by Non-Veg.
- Cuisine preferences: a wide variety, but sparsely populated.
- Numeric variables (BMI, spice tolerance, sweet tooth): some separation across

6. MODEL & METHODOLOGY

6.1 Logistic Regression (Baseline)

- Served as the baseline model.
- Performed poorly due to imbalance.
- Produced warnings: some classes had **0 precision/recall** (no predictions).
- Report note: "Our baseline Logistic Regression achieved low macro F1 and failed to predict minority classes (UndefinedMetricWarnings occurred)."

6.2 Random Forest Classifier (Final Model)

- Selected for its ability to handle non-linear relationships.
- Tuned with **GridSearchCV** (3-fold).
- Best hyperparameters:
  - n_estimators = 100
  - max_depth = None
  - min_samples_leaf = 1
- Provided much better performance and explainability.

7. EVALUATION

Metrics used:

- Accuracy
- Precision, Recall, and F1-score
- Macro F1 (important due to imbalance)
- Confusion Matrix

Key Findings:

- Non-Veg: predicted very accurately (19/20 correct).
- Minority classes: consistently misclassified as Non-Veg.
- Macro F1: low, due to poor performance on minority classes.

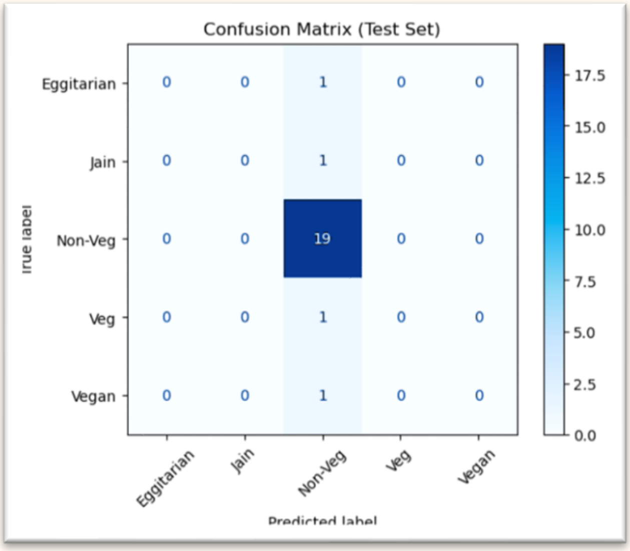

Fig.2 : Showing Model predictions vs true lebels

8. EXPLAINABILITY & INSIGHTS

Random Forest feature importances revealed which features were most influential.

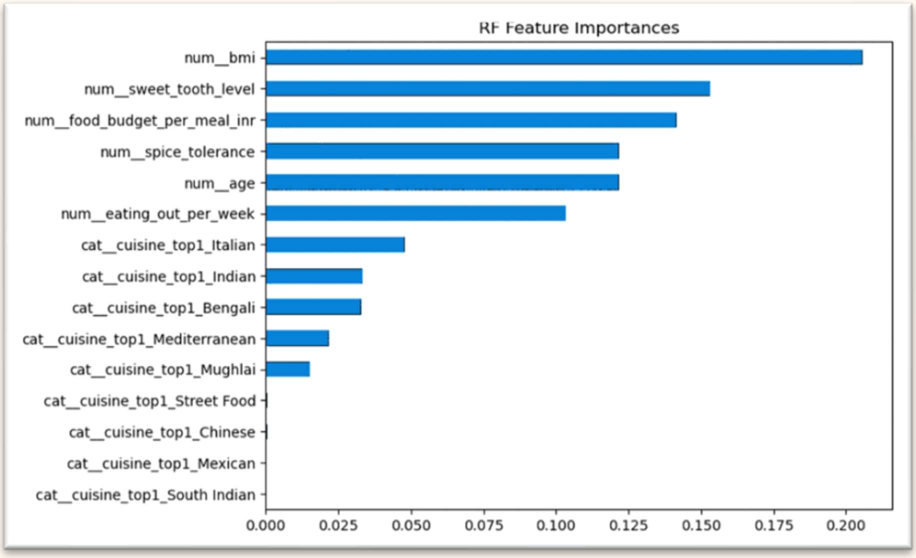

Fig. 3 : Random Forest Feature Importance for predicting dietary preference

Key Insights:

- BMI and sweet-tooth level: strongest predictors.
- Food budget per meal and spice tolerance: secondary predictors.
- Cuisine preferences (Italian, Indian, Bengali): moderate influence.
- Other cuisines: little effect.

This suggests that health metrics and taste preferences strongly align with dietary choices, guiding canteen menu planning.

9. RESULTS & DISCUSSION

- **Best model:** Random Forest.
- **Performance:** High accuracy for Non-Veg, poor generalization for minority classes.
- **Insights:**
  - Stocking Non-Veg items is justified (majority demand).
  - Minority preferences are underrepresented, requiring more balanced data.

10. LIMITATIONS

1) **Small Dataset** : Only 111 Rows
2) **Severe imbalance**: minority classes <5 samples.
3) Cross-validation produced warnings due to **tiny classes**.
4) Logistic Regression ineffective.
5) Random Forest biased toward majority class.

11. FUTURE WORK

1) Collect **larger, balanced datasets**.
2) Apply resampling (SMOTE, class weights).
3) Explore advanced models (XGBoost, LightGBM, Neural Networks).
4) Add lifestyle features (e.g., exercise, eating-out frequency).
5) Deploy as a **canteen decision-support web app**.

12. SOFTWARE REQUIREMENTS & TOOLS

- **Programming Language**
  - Python 3.9+

- **Libraries & Frameworks**
  - scikit-learn (for model training & evaluation)
  - imbalanced-learn (for handling imbalance, SMOTE)
  - pandas, NumPy (data handling & preprocessing)
  - Matplotlib, Seaborn (visualization)

- **Development Environment**
  - Jupyter Notebook (Anaconda Distribution)

- **Version Control**
  - Git & GitHub (for code hosting and collaboration)

- **Hardware**
  - Standard laptop/desktop with at least 4 GB RAM (no GPU required for this project).

13. CONCLUSION

The Canteen Menu Optimizer demonstrates how machine learning can support canteen operations.

- Logistic Regression struggled; Random Forest performed better.
- Non-Veg was predicted accurately, but minority classes failed.
- Feature analysis showed BMI, sweet-tooth level, spice tolerance, and cuisines were strong predictors.

While limited by dataset size and imbalance, the project provides useful insights and lays the foundation for future improvements.

14. REFERENCES

- ✓ Pedregosa et al. (2011). *Scikit-learn: Machine Learning in Python.*
- ✓ scikit-learn documentation. https://scikit-learn.org/
- ✓ imbalanced-learn documentation. https://imbalanced-learn.org/
- ✓ Matplotlib & Seaborn documentation.
- ✓ OpenAI. (2025). ChatGPT (GPT-5) [Large language model]. Retrieved September 2025, from https://chat.openai.com