Supervised Machine Learning Final Project

Title: *Using Classification Models to Guide Dietary Choices for Diabetics from Real-World Food Data*

1. Objective of the Analysis

The primary objective of this analysis is to develop a supervised machine learning model that helps classify food items based on their nutritional attributes into one of three categories for diabetic individuals: **More Often**, **In Moderation**, or **Less Often**. This classification provides a data-driven guideline that can support patients and healthcare professionals in making informed dietary decisions.

The focus of this analysis is **predictive modelling**, with a secondary emphasis on **interpretability** to ensure transparency around which nutritional features most influence the classification outcomes. By understanding the relationships between nutrients and dietary recommendations, stakeholders such as nutritionists, healthcare providers, and diabetic patients themselves can benefit from actionable insights.

This model could serve as the basis for a nutrition recommendation engine integrated into health apps or meal planning tools designed for diabetic populations.

2. Dataset Description

The dataset used in this analysis is a real-world nutritional dataset containing detailed information on various food items, with the goal of supporting dietary recommendations for

individuals with diabetes. Each row represents a food item, and each column provides a nutritional attribute relevant to diabetic health management.

Key features of the dataset include:

- **Energy (kcal)** – Caloric content
- **Carbohydrates (g)** – Total carbohydrates
- **Sugars (g)** – Including added sugars
- **Total Fat (g)**, **Saturated Fat (g)** – Important for cardiovascular health
- **Fiber (g)** – Aiding blood sugar control
- **Protein (g)** – Helps with satiety and glycemic control
- **Sodium (mg)** – Linked to hypertension risk, common in diabetics
- **Cholesterol (mg)** – Important in diabetic heart disease risk
- **Calcium, Iron, Potassium** – Essential micronutrients

The target variable (class) indicates whether the food should be consumed:

- **More Often** – Healthier choices, lower in sugar, fat, and sodium
- **In Moderation** – Acceptable if portion-controlled
- **Less Often** – High in sugar, fat, sodium or calories, and potentially harmful in large quantities

The dataset was preprocessed to remove irrelevant features, handle missing values, and prepare it for classification modeling. Feature scaling was applied where appropriate, and class distribution was checked to ensure balanced model training.

	Calories	Total Fat	Saturated Fat	Monounsaturated Fat	Polyunsaturated Fat	Trans Fat	Cholesterol	Sodium	Total Carbohydrate	Dietary Fiber	Sugars	Sugar Alcohol	Protein	Vitamin A	Vitamin C	Calcium	Iron	class
0	149.0	0	0.0	0.0	0.0	0.0	0	9.0	9.8	0.0	0.0	0	1.3	0	0	0	0	'In Moderation'
1	123.0	0	0.0	0.0	0.0	0.0	0	5.0	6.6	0.0	0.0	0	0.8	0	0	0	0	'In Moderation'
2	150.0	0	0.0	0.0	0.0	0.0	0	4.0	11.4	0.0	0.0	0	1.3	0	0	0	0	'In Moderation'
3	110.0	0	0.0	0.0	0.0	0.0	0	6.0	7.0	0.0	0.0	0	0.8	0	0	0	0	'In Moderation'
4	143.0	0	0.0	0.0	0.0	0.0	0	7.0	13.1	0.0	0.0	0	1.0	0	0	0	0	'In Moderation'
5	110.0	0	0.0	0.0	0.0	0.0	0	6.0	7.0	0.0	0.0	0	0.8	0	0	0	0	'In Moderation'

Figure 1. The first six row of the dataset

3. Data Exploration, Cleaning, and Feature Engineering

Prior to training any models, the dataset underwent exploratory analysis and preprocessing to ensure it was clean, consistent, and suitable for classification.

Key steps taken:

- **Missing Value Handling:** Columns with missing or irrelevant values were removed. The remaining data was checked for nulls and cleaned accordingly.

- **Feature Selection:** Nutritionally relevant attributes such as energy, sugars, carbohydrates, fats, sodium, fiber, and protein were retained, while non-nutritive or redundant fields were dropped.
- **Class Balance Check:** The target variable (class) was assessed for balance across the three categories (More Often, In Moderation, Less Often). Class representation was adequate to proceed without resampling.
- **Standardization:** Numerical features were standardized using StandardScaler to ensure fair treatment across models that are sensitive to feature scaling (e.g., Logistic Regression, SVM).
- **Train-Test Split:** The data was split into training and testing subsets (80/20) to evaluate model generalization.

To better understand relationships between features, we generated a **correlation heatmap**, as shown below.

Figure 1. Correlation Heatmap of Nutritional Attributes

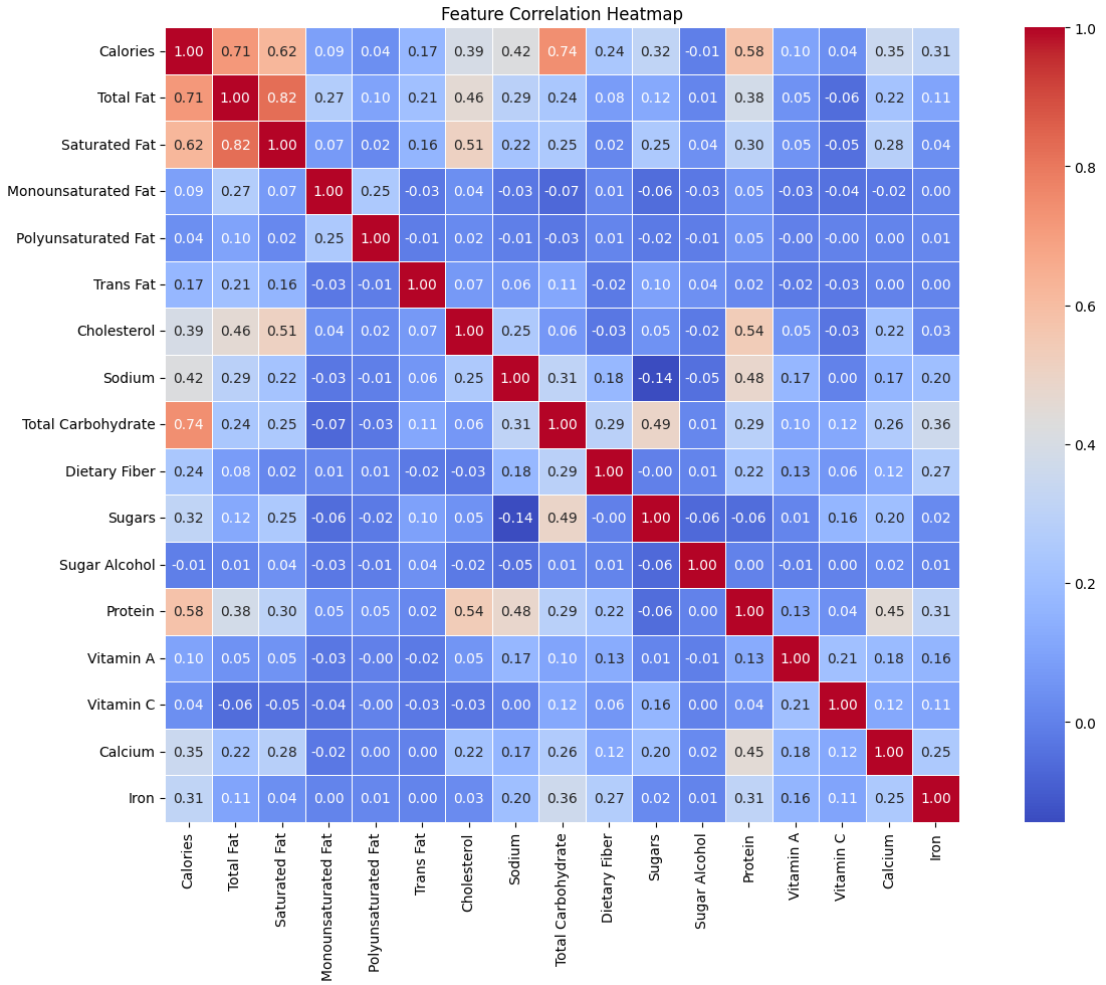

Figure 2. Correlation Heatmap of Nutritional Attributes

This visualization reveals several moderate to strong correlations — for example, calories and carbohydrates, or total fat and saturated fat. Recognizing these relationships helps

interpret model behavior and identify potential collinearity issues, particularly relevant for linear models like logistic regression.

4. Classifier Models Trained

To achieve our objective of classifying food items into diabetic dietary categories, we trained and evaluated three supervised classification models of varying complexity and interpretability:

1. **Logistic Regression**
    - A baseline linear classifier for interpretation.
    - Offers insight into the importance of features through model coefficients.
    - Performs well when classes are linearly separable.
2. **Support Vector Machine (SVC)**
    - Kernel-based model for higher-dimensional classification.
    - Effective for small to medium-sized datasets.
    - Can find non-linear boundaries.
3. **Random Forest Classifier**
    - Ensemble-based model using decision trees.
    - Provides high accuracy and feature importance rankings.
    - Handles non-linearities and interactions between features well.

All models were trained using the same standardized features and evaluated on the same test set to ensure consistency in performance comparison. Each model's predictions were assessed using key metrics: **accuracy**, **precision**, **recall**, and **F1-score**.

A **confusion matrix** was also generated for each model to visualize prediction correctness across the three classes. These matrices helped assess where models performed best, and where misclassifications occurred (e.g., confusing "In Moderation" with "Less Often").

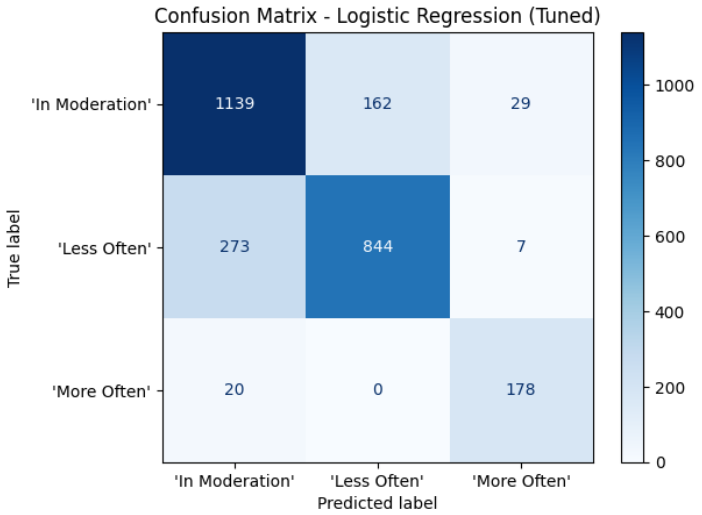

Figure 3. Confusion Matrix of Logistic Regression

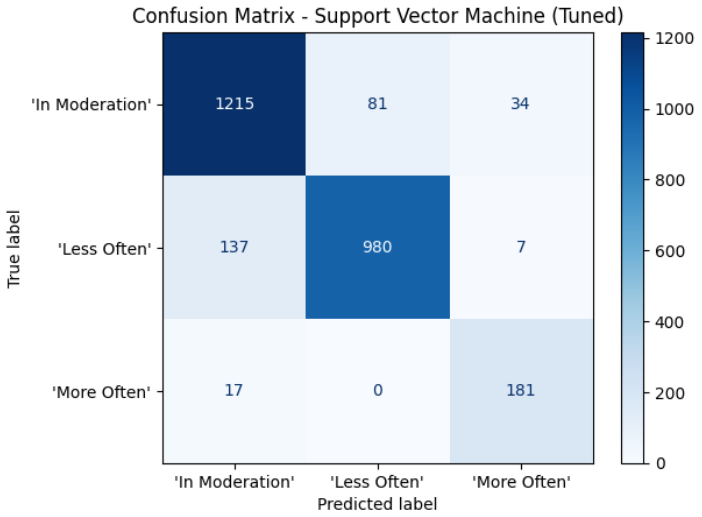

Figure 4. Confusion Matrix of Support Vector Machine

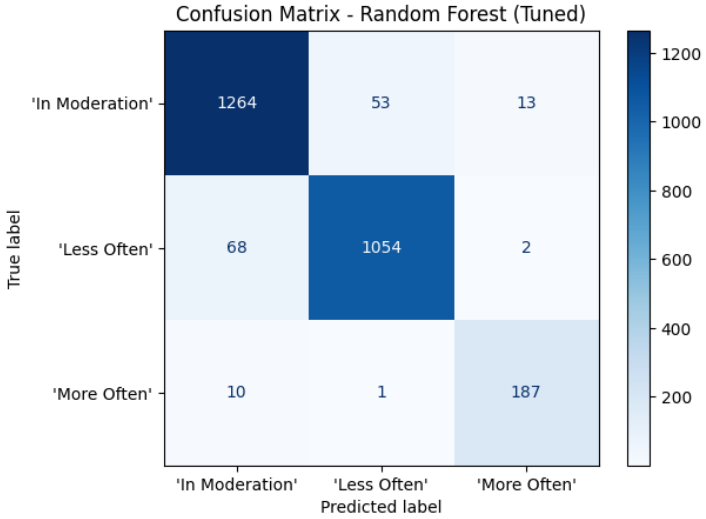

Figure 5. Confusion Matrix of Random Forest

To summarize the model performances across metrics, we created the following comparison chart:

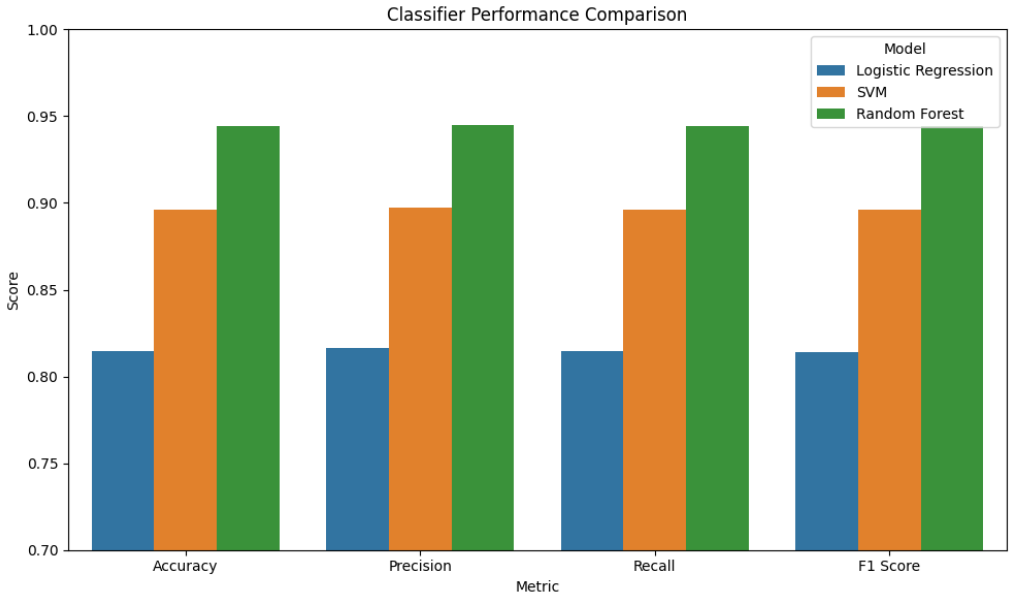

Figure 6. Classifier Performance Comparison

This visualization clearly shows that the **Random Forest** model outperformed the others in most evaluation metrics, while maintaining acceptable interpretability. Its balance of predictive power and insight into feature importance made it the ideal choice for this use case.

5. Final Model Selection

After evaluating all three classifiers: Logistic Regression, Support Vector Machine (SVM), and Random Forest, the **Random Forest Classifier** was selected as the final model for this analysis.

This choice was based on the following considerations:

- **Performance:**
  Random Forest achieved the highest scores across all evaluation metrics, particularly **accuracy**, **precision**, and **F1-score**, indicating strong overall predictive performance.
- **Handling of Nonlinearity:**
  Compared to Logistic Regression, Random Forest effectively captured complex interactions and nonlinear relationships between nutritional features and class labels.
- **Interpretability through Feature Importance:**
  Despite being an ensemble model, Random Forest provides an intuitive ranking of feature importance, which aligns well with the health-centered goal of this project. Understanding which nutrients most influence recommendations enhances stakeholder trust and decision-making transparency.

- **Robustness:**
  Random Forest is resilient to outliers and overfitting due to its ensemble nature, making it a reliable choice for real-world health datasets.

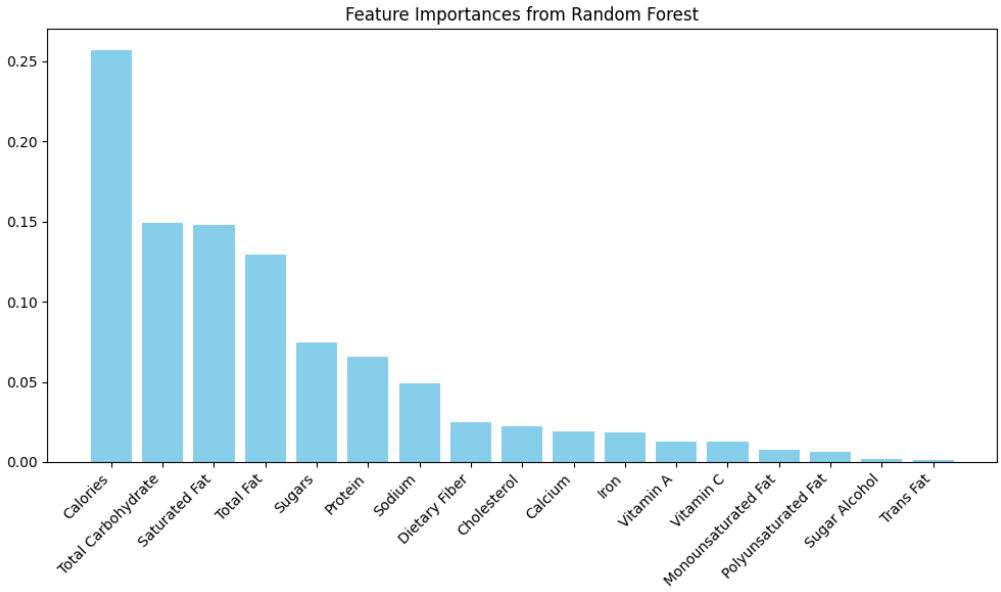

Figure 7. Feature Importance Plot from Random Forest

This figure highlights the most influential features in determining whether a food item should be consumed more often, in moderation, or less often. Nutrients such as **sugar**, **sodium**, **saturated fat**, and **total calories** emerged as top contributors — aligning with clinical dietary guidelines for diabetic individuals.

6. Key Findings and Insights

The classification modeling and data exploration yielded several important insights relevant to dietary guidance for diabetic individuals:

Nutritional Drivers of Classification
Based on the Random Forest model's feature importance analysis, the top factors influencing food classification were:

- **Sugar (g):** Most predictive of "Less Often" classification, which aligns with dietary restrictions for diabetic individuals.
- **Sodium (mg):** Foods with high sodium levels tended to be classified as "Less Often" or "In Moderation," reflecting cardiovascular health concerns.
- **Saturated Fat (g):** Heavily influenced negative classifications due to its association with increased health risks in diabetic patients.
- **Energy (kcal):** High caloric content was a common trait among foods classified as "Less Often."
- **Fiber (g):** Associated with "More Often" classification, consistent with its role in stabilizing blood sugar levels.

- **Carbohydrates (g):** A contributing factor whose impact varied depending on sugar content and food context.

Model Performance Insights

- The **Random Forest** model achieved the best overall performance across all metrics, especially in differentiating between the "More Often" and "Less Often" categories.
- **Support Vector Machine (SVM)** showed moderate performance but was less accurate in identifying the "In Moderation" class.
- **Logistic Regression**, while interpretable, struggled with the multiclass classification task, particularly in capturing nonlinear feature interactions.

Confusion matrix analyses revealed some misclassification between the "In Moderation" class and the other two categories, which may reflect the nuanced nature of dietary decisions where portion size and food context play a role.

Alignment with Health Guidelines
The model's behavior aligns closely with standard dietary recommendations for diabetes management:

- Foods low in **sugar**, **sodium**, and **saturated fat** and high in **fiber** are promoted for more frequent consumption.
- Nutrient-dense items with moderate calories are more likely to be recommended as healthy options.

These insights demonstrate the model's potential as a practical tool for supporting diabetes-friendly dietary planning. Its interpretability and accuracy make it suitable for informing patient decisions or powering recommendation features in digital health applications.

7. Limitations and Future Steps

While the analysis produced valuable insights and a high-performing model, there are several limitations to consider, along with opportunities for future improvement.

Limitations

- **Limited Feature Granularity:**
  The dataset, while rich in nutritional content, does not include contextual information such as serving size variations, food preparation methods, or brand-specific differences, which can affect health recommendations.
- **No Personalization:**
  The model does not account for individual patient needs, preferences, or conditions (e.g., age, activity level, insulin sensitivity), which limits its use in personalized diet planning.
- **Class Overlap and Ambiguity:**
  Some food items naturally fall between categories (e.g., a moderately salty snack with

high fiber), leading to classification ambiguity. This is reflected in the confusion between the “In Moderation” class and the others.

- **Model Generalization Risk:**
  Although the model performed well on the current dataset, it has not been tested on new or external food items, and may require retraining or validation before deployment in real-world tools.

Next Steps

- **Integrate Additional Features:**
  Enrich the dataset with contextual attributes such as portion size, glycemic index, food groups, or preparation methods. This could improve both accuracy and real-world applicability.
- **Explore Personalization Models:**
  Future versions of this project could include user-specific inputs and use multi-input models or recommendation systems tailored to individual dietary needs.
- **Deploy and Validate in Real-World Settings:**
  Testing the model within an app or a clinical decision support system with feedback loops from real users could validate and improve its utility.
- **Compare with Additional Models:**
  Incorporating other classifiers such as Gradient Boosting Machines or deep learning models may yield even better performance, especially with richer datasets.