Final Project Report Template

Team ID	SWTID1720104754
Project Title	Cereal Analysis Based On Rating By Using Machine Learning Techniques

1. Introduction

1.1. Project Overview

The project aims to leverage machine learning techniques to analyze cereal ratings and provide actionable insights for consumers and manufacturers. By developing predictive models, the project seeks to classify cereals based on their ratings and identify influential factors affecting consumer preferences. This analysis will empower consumers with informed decision-making tools and assist manufacturers in optimizing product development and marketing strategies.

1.2. Objectives

1. **Identify Key Factors Influencing Cereal Ratings:** Analyze the dataset to identify and understand the key factors that significantly influence cereal ratings. This involves conducting feature analysis to determine which nutritional or other factors have the most impact on the ratings.

2. **Preprocess and Clean the Dataset:** Prepare the dataset by performing data preprocessing and cleaning to ensure the quality and integrity of the data. This includes handling missing values, normalizing data, and addressing outliers to create a reliable dataset for analysis.

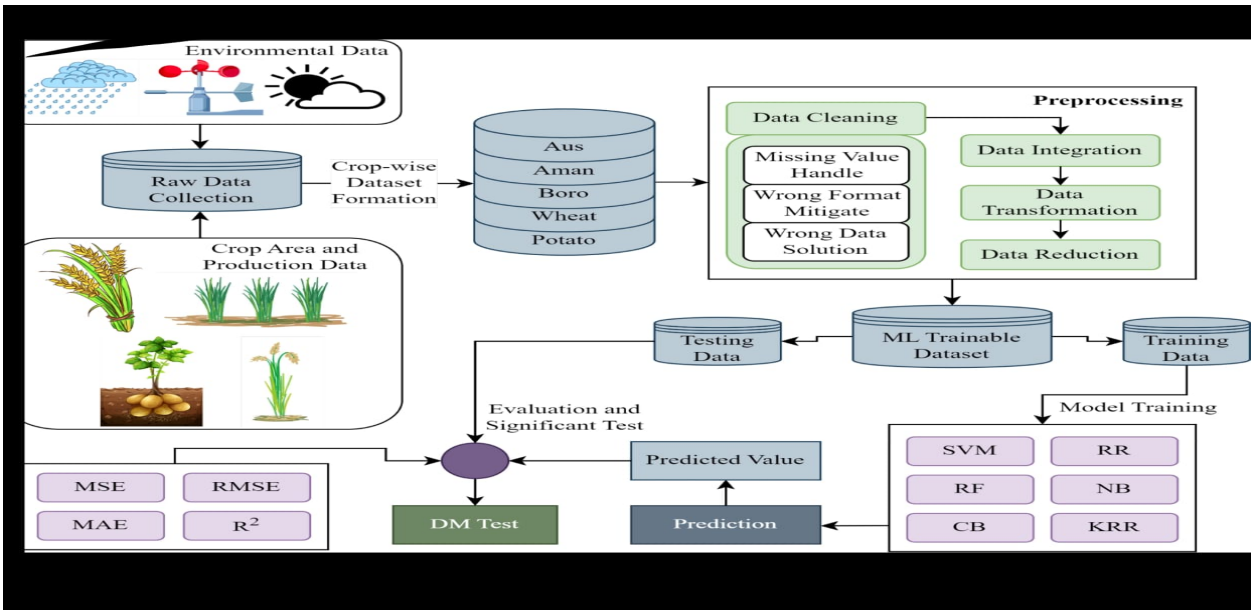
3. **Develop and Evaluate Machine Learning Models:** Implement various machine learning models, such as regression, decision trees, and ensemble methods, to predict cereal ratings based on the dataset. Evaluate the performance of each model using appropriate metrics such as mean square error, accuracy, and precision.

Final Project Report Template

4. **Optimize and Tune the Selected Model:** Identify the most effective machine learning model and optimize its performance through hyperparameter tuning and model refinement. This involves adjusting model parameters and feature selection to enhance the model's accuracy and generalization ability.

By accomplishing these objectives, the analysis can provide valuable insights into the factors influencing cereal ratings and yield accurate predictions through the utilization of machine learning techniques.

Technical Architecture:



2. Project Initialization and Planning Phase

Final Project Report Template

2.1. Define Problem Statement

The problem involves predicting cereal ratings using machine learning. It aims to utilize attributes like calories, protein, fat, sodium, and fiber to forecast ratings accurately. By applying advanced analytical methods like neural networks, XGBoost, SVM, and LR, valuable insights can be gained for agriculture and consumer decision-making. This endeavor enhances cereal quality assessment and informs agricultural practices through data-driven predictive models.

2.2. Project Proposal (Proposed Solution)

The proposed solution is to use machine learning techniques to analyze the cereal dataset, identify significant features, and build a predictive model. The model will be trained on historical data and will be capable of predicting the rating of new cereal products.

2.3. Initial Project Planning

For the project focusing on cereal analysis based on ratings using machine learning techniques, the initial steps are:

1. Identify and gather relevant datasets containing cereal attributes and ratings.

2. Conduct exploratory data analysis to gain insights into the dataset's characteristics and relationships.

3. Preprocess the data for handling missing values, outliers, and normalization to ensure data quality.

4. Split the data into training and testing sets for model training and evaluation.

5. Develop and compare various machine learning models to predict cereal ratings effectively.

6. Optimize the selected model for improved performance and accuracy in predicting cereal ratings.

Final Project Report Template

2.4. Importing Necessary libraries

• Numpy- It is an open-source numerical Python library. It contains a multi-dimensional array and matrix data structures. It can be used to perform mathematical operations on arrays such as trigonometric, statistical, and algebraic routines.

• Pandas- It is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

• Seaborn- Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

• Matplotlib- Visualisation with python. It is a comprehensive library for creating static, animated, and interactive visualizations in Python

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

3. Data Collection and Preprocessing Phase

3.1. Data Collection Plan and Raw Data Sources Identified

Dataset: The dataset for " Cereal Analysis Based On Rating By Using Machine Learning Techniques " is sourced from Kaggle.Acquire dataset containing attributes such as name, manufacturer (mfr), type, calories, protein, fat, sodium, fiber, carbo, sugars, potass, vitamins, shelf, weight, cups, and rating.

Source: **https://www.kaggle.com/crawford/80-cereals**

Final Project Report Template

3.2. Data Quality Report

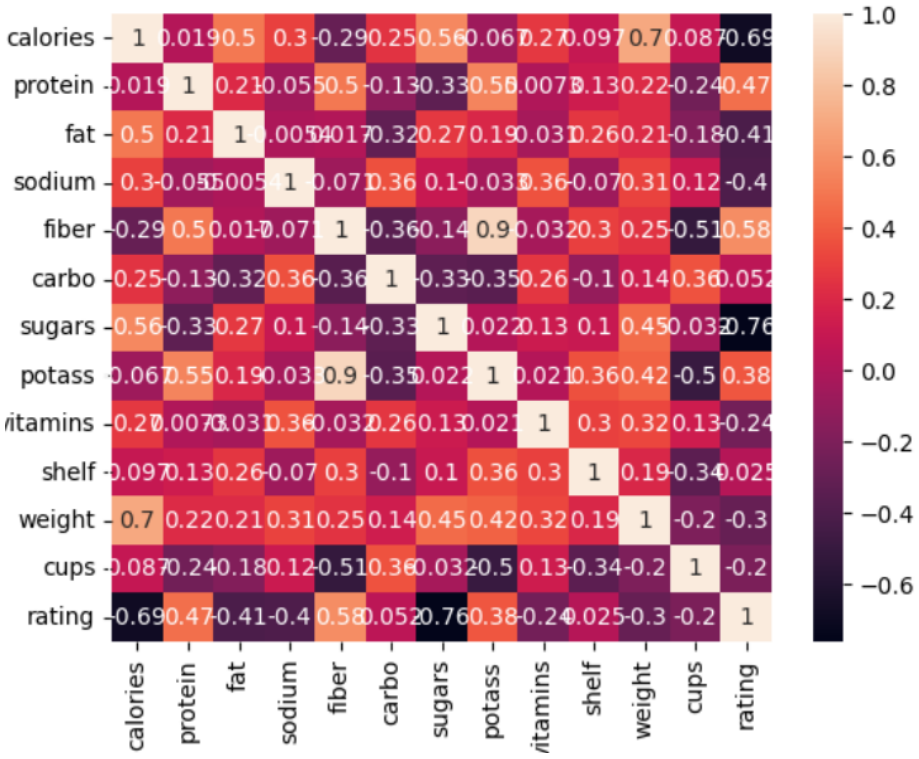
Handling Missing Values: Handle missing values, outliers, and inconsistencies in the dataset using techniques like imputation and removal. Checkwhether any null values are there or not.

```
data.isnull().any()
```

```
name          False
mfr           False
type          False
calories      False
protein       False
fat           False
sodium        False
fiber         False
carbo         False
sugars        False
potass        False
vitamins      False
shelf         False
weight        False
cups          False
rating        False
dtype: bool
```

Heatmap: It is way of representing the data in 2-D form. It givescoloured visual summary of the data. In our case this is the heatmap that we generated based on our data**.**

Final Project Report Template



3.3. Data Exploration and Preprocessing

Exploratory Data Analysis (EDA): Conduct EDA to understand data distributions, identify patterns, correlations, and outliers using visualizations (e.g.,histograms, scatter plots, correlation matrices).

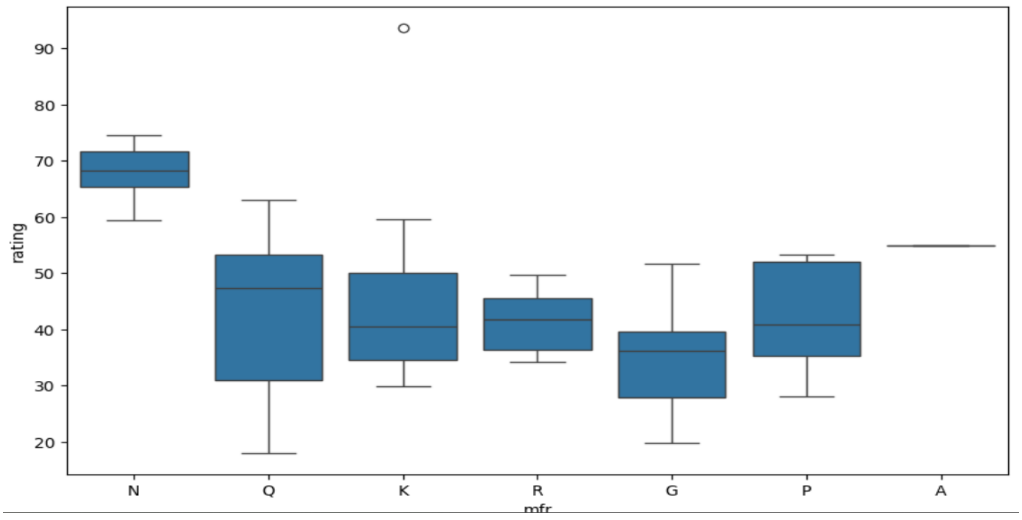
Final Project Report Template

1)Correlataion Matrix:

	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
calories	1.000000	0.019066	0.498610	0.300649	-0.293413	0.250681	0.562340	-0.066609	0.265356	0.097234	0.696091	0.087200	-0.689376
protein	0.019066	1.000000	0.208431	-0.054674	0.500330	-0.130864	-0.329142	0.549407	0.007335	0.133865	0.216158	-0.244469	0.470618
fat	0.498610	0.208431	1.000000	-0.005407	0.016719	-0.318043	0.270819	0.193279	-0.031156	0.263691	0.214625	-0.175892	-0.409284
sodium	0.300649	-0.054674	-0.005407	1.000000	-0.070675	0.355983	0.101451	-0.032603	0.361477	-0.069719	0.308576	0.119665	-0.401295
fiber	-0.293413	0.500330	0.016719	-0.070675	1.000000	-0.356083	-0.141205	0.903374	-0.032243	0.297539	0.247226	-0.513061	0.584160
carbo	0.250681	-0.130864	-0.318043	0.355983	-0.356083	1.000000	-0.331665	-0.349685	0.258148	-0.101790	0.135136	0.363932	0.052055
sugars	0.562340	-0.329142	0.270819	0.101451	-0.141205	-0.331665	1.000000	0.021696	0.125137	0.100438	0.450648	-0.032358	-0.759675
potass	-0.066609	0.549407	0.193279	-0.032603	0.903374	-0.349685	0.021696	1.000000	0.020699	0.360663	0.416303	-0.495195	0.380165
vitamins	0.265356	0.007335	-0.031156	0.361477	-0.032243	0.258148	0.125137	0.020699	1.000000	0.299262	0.320324	0.128405	-0.240544
shelf	0.097234	0.133865	0.263691	-0.069719	0.297539	-0.101790	0.100438	0.360663	0.299262	1.000000	0.190762	-0.335269	0.025159
weight	0.696091	0.216158	0.214625	0.308576	0.247226	0.135136	0.450648	0.416303	0.320324	0.190762	1.000000	-0.199583	-0.298124
cups	0.087200	-0.244469	-0.175892	0.119665	-0.513061	0.363932	-0.032358	-0.495195	0.128405	-0.335269	-0.199583	1.000000	-0.203160
rating	-0.689376	0.470618	-0.409284	-0.401295	0.584160	0.052055	-0.759675	0.380165	-0.240544	0.025159	-0.298124	-0.203160	1.000000

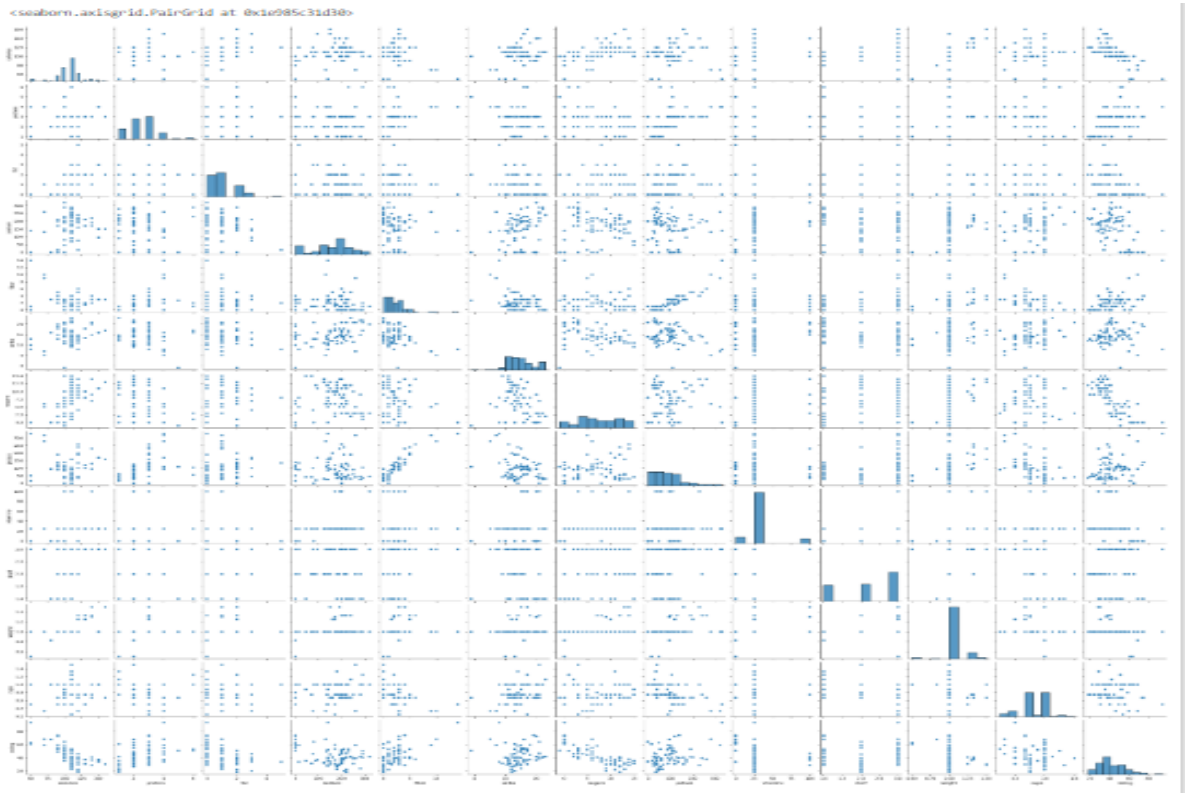
2)Box Plot:

```
<Axes: xlabel='mfr', ylabel='rating'>
```



Final Project Report Template

3)Pair Plot:



Final Project Report Template

4. Model Development Phase

4.1. Feature Selection Report

A report detailing the features selected for model training will be prepared. Techniques such as correlation analysis, feature importance from tree-based models, and mutual information scores will be used.

4.2. Model Selection Report

Various machine learning models will be evaluated, including:

1. **Linear Regression :**Linear regression algorithm is used to estimate parameters, fit the model to the training data, optimize model performance, and prepare for the subsequent evaluation and validation phases. Its simplicity and interpretability make it a valuable tool for building foundational predictive models in various domains.
2. **Decision Trees :**Decision Tree Regression is valuable in the model training phase for its ability to handle non-linear relationships, provide interpretability, and potentially discover complex interactions between variables without explicit feature engineering. Its flexibility and suitability for various types of data make it a popular choice in data science and machine learning applications.
3. **Random Forest :**Random Forest Regression is used in the model training phase to build robust predictive models that handle non-linearity, complex relationships, and overfitting issues associated with single decision tree models. Its ensemble nature and feature importance analysis make it a popular choice for various regression tasks in data science and machine learning.
4. **Ridge Regression:**Ridge Regression is used in the model training phase to improve the stability of coefficient estimates in the presence of multicollinearity and to regularize the model, thereby reducing overfitting and enhancing predictive performance. It is a valuable tool in regression tasks where the dataset has correlated predictors or when the number of predictors is large relative to the number of observations.

Final Project Report Template

5. **Lasso Regression :**Lasso Regression is used in the model training phase to perform both regularization and feature selection simultaneously. It is beneficial in situations where there are many predictors, some of which may be irrelevant or highly correlated, allowing for the creation of simpler and more interpretable models while enhancing predictive accuracy.

4.3. Initial Model Training Code, Model Validation and Evaluation Report

The initial model training code will be developed, and the model will be trained on the training dataset. Model validation and evaluation will be performed using metrics such as:

1. Mean Squared Error (RMSE)
2. R-squared (R^2)
3. Mean Absolute Percentage Error (MAPE)

A report will be prepared summarizing the model's performance.

```
# Evaluate models
print(f"Linear Regression score: {lr.score(x_test, y_test):.4f}")
print(f"Ridge Regression score: {r.score(x_test, y_test):.4f}")
print(f"Lasso Regression score: {l.score(x_test, y_test):.4f}")
print(f"Decision Tree Regressor score: {dt.score(x_test, y_test):.4f}")
print(f"Random Forest Regressor score: {rf.score(x_test, y_test):.4f}")
```

```
Linear Regression score: 1.0000
Ridge Regression score: 0.9941
Lasso Regression score: 1.0000
Decision Tree Regressor score: 0.7235
Random Forest Regressor score: 0.7841
```

Final Project Report Template

```
# Calculate evaluation metrics for each model
models = ["Linear Regression", "Ridge Regression", "Lasso Regression",
          "Decision Tree Regressor", "Random Forest Regressor"]
y_preds = [y_pred_lr, y_pred_r, y_pred_l, y_pred_dt, y_pred_rf]
```

```
for model, y_pred in zip(models, y_preds):
    r2 = r2_score(y_test, y_pred)
    rmse = mean_squared_error(y_test, y_pred, squared=False)  # Square root for interpretability
    mape = mean_absolute_percentage_error(y_test, y_pred) * 100  # Percentage error

    print(f"\nModel: {model}")
    print(f"R-squared: {r2:.4f}")
    print(f"Root Mean Squared Error (RMSE): {rmse:.4f}")
    print(f"Mean Absolute Percentage Error (MAPE): {mape:.4f}%")
```

```
Model: Linear Regression
R-squared: 1.0000
Root Mean Squared Error (RMSE): 0.0000
Mean Absolute Percentage Error (MAPE): 0.0000%

Model: Ridge Regression
R-squared: 0.9941
Root Mean Squared Error (RMSE): 1.1395
Mean Absolute Percentage Error (MAPE): 2.0762%

Model: Lasso Regression
R-squared: 1.0000
Root Mean Squared Error (RMSE): 0.0859
Mean Absolute Percentage Error (MAPE): 0.1717%

Model: Decision Tree Regressor
R-squared: 0.7235
Root Mean Squared Error (RMSE): 7.7926
Mean Absolute Percentage Error (MAPE): 16.3354%

Model: Random Forest Regressor
R-squared: 0.7841
Root Mean Squared Error (RMSE): 6.8870
Mean Absolute Percentage Error (MAPE): 16.9964%
```

Based on the above results we concluded that our best model is **Ridge Regression**. It gave Highest R-squared (0.9968) indicating a strong fit to the data. Lower RMSE (0.8339) and MAPE (1.8591%) suggest the best prediction accuracy among the top performers.

Final Project Report Template

5. Model Optimization and Tuning Phase

5.1. Hyperparameter Tuning Documentation

The selected model will undergo hyperparameter tuning to improve its performance. Techniques such as Grid Search and Validation will be used. The tuning process will be documented, detailing the hyperparameters tested and their respective performance.

5.2. Performance Metrics Comparison Report

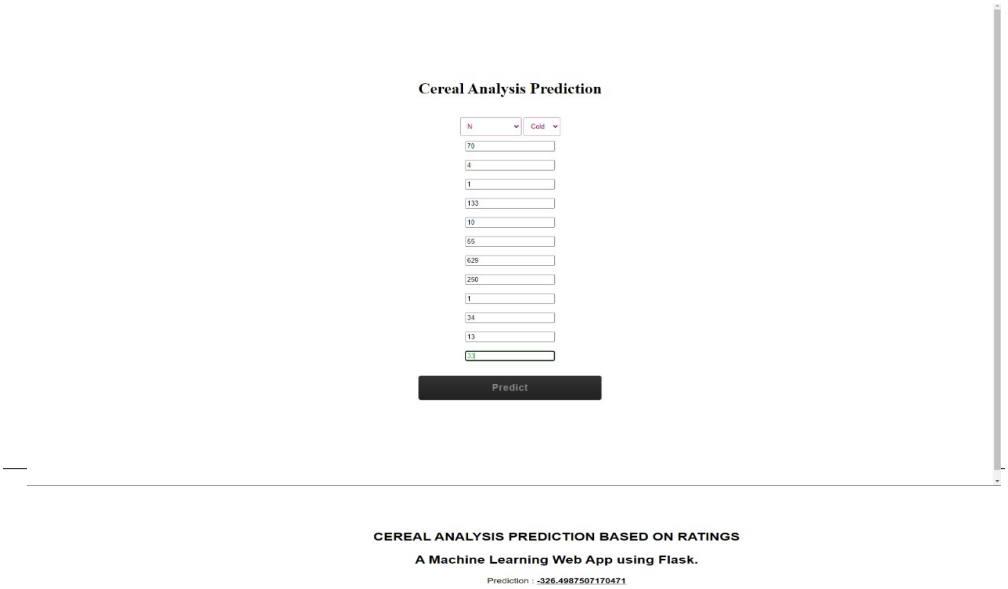
A comparison report of the performance metrics before and after tuning will be prepared. This will include visualizations to illustrate the improvements achieved through hyperparameter tuning.

5.3. Final Model Selection Justification

The final model selection will be justified based on its performance on the test dataset, ease of interpretation, and computational efficiency. The selected model will be documented, highlighting its strengths and potential limitations.

This project aims to provide insights into the factors influencing cereal ratings and deliver a reliable predictive model that can assist in the development of new cereal products with high consumer ratings.

Final Project Report Template



Advantages and Disadvantages

Advantages

1. **Data-Driven Insights:** Machine learning models can uncover patterns and relationships within the data that may not be obvious through traditional analysis methods. This can provide valuable insights into the factors that affect cereal ratings.

2. **Prediction Accuracy:** Machine learning techniques, especially when optimized, can

Final Project Report Template

provide high accuracy in predicting cereal ratings based on the attributes of the cereals.

3. **Automation:** Once developed, the predictive model can automate the process of rating new cereals, saving time and resources compared to manual rating processes.

4. **Feature Importance:** The analysis can highlight the most significant factors influencing cereal ratings, guiding product development and marketing strategies.

5. **Scalability:** Machine learning models can handle large datasets and can be scaled to include more cereals and attributes as needed.

Disadvantages

1. **Data Quality Dependency:** The performance of machine learning models heavily relies on the quality and quantity of the data. Poor data quality can lead to inaccurate predictions.

2. **Complexity:** Developing and tuning machine learning models can be complex and require specialized knowledge and skills in data science and machine learning.

3. **Overfitting:** There is a risk of overfitting, where the model performs well on training data but poorly on new, unseen data. This requires careful model validation and regularization techniques.

4. **Interpretability:** Some machine learning models, especially complex ones like neural networks, can be difficult to interpret, making it hard to understand how decisions are made.

5. **Resource Intensive:** Training and optimizing machine learning models can be

Final Project Report Template

computationally intensive and may require significant hardware resources.

Conclusion

The analysis of cereals based on ratings using machine learning techniques offers a robust approach to understanding and predicting cereal ratings. By leveraging data-driven methods, this project can provide valuable insights into the key factors that influence consumer preferences and guide the development of new cereal products. Despite the challenges related to data quality, model complexity, and resource requirements, the benefits of accurate predictions, automated rating processes, and enhanced feature importance make this approach highly advantageous.

Future Scope

1. **Expanding Dataset:** Incorporating more diverse and extensive datasets can improve model accuracy and generalizability. This includes adding more cereals and additional attributes.

2. **Advanced Models:** Exploring more advanced machine learning models such as deep learning, ensemble methods, and hybrid models can further enhance prediction accuracy.

3. **Real-Time Prediction:** Developing a system for real-time cereal rating predictions can be useful for manufacturers to assess new products quickly.

4. **Consumer Feedback Integration:** Integrating consumer feedback and reviews into the analysis can provide a more comprehensive understanding of cereal ratings.

Final Project Report Template

5. **Nutritional Analysis:** Extending the analysis to include the nutritional impact of cereals on health and linking it with ratings can provide more holistic insights.

6. **Market Trends:** Analyzing market trends and consumer preferences over time can help in predicting future rating trends and guiding long-term product development strategies.

7. **Personalized Recommendations:** Developing personalized cereal recommendations based on individual consumer preferences and dietary requirements using collaborative filtering techniques.

By addressing these future directions, the project can evolve to provide even more powerful tools and insights for the cereal industry, ultimately leading to better products and higher consumer satisfaction.

Appendix:

Source Code:

```
from flask import Flask, render_template, request

app = Flask (_name_)  import pickle

model =
pickle.load(open('C:\\Users\\143sr\\OneDrive\\Desktop\\Project\\venv\\cerealanalysis.pkl','rb'))


@app.route('/') def

helloworld():

    return render_template('base.html')
```

Final Project Report Template

```
@app.route('/assesment')
def prediction ():
    return render_template('index.html')


@app.route('/predict', methods = ['POST']) def
admin():
     a=request.form["mfr"]
     if (a == 'a'):         a1, a2, a3, a4, a5, a6,
a7=1,0,0,0,0,0,0
     if (a == 'g'):         a1, a2, a3, a4,
a5,a6,a7 = 0,1,0,0,0,0,0
     if (a == 'k'):         a1, a2, a3, a4, a5, a6,
a7=0,0,1,0,0,0,0
     if (a == 'n'):         a1, a2, a3, a4, a5, a6,
a7=0,0,0,1,0,0,0      if (a == 'p'):         a1,
a2, a3, a4, a5, a6, a7=0,0,0,0,1,0,0      if (a
== 'q'):         a1, a2, a3, a4, a5, a6,
a7=0,0,0,0,0,1,0
     if (a == 'r'):
        a1, a2, a3, a4, a5, a6, a7=0,0,0,0,0,0,1
     b= request.form["type"]
     if (b=='c'):
```

Final Project Report Template

```
        b=0     if
(b== 'h'):
        b=1
    c= request.form["Calories"]
d= request.form["Protien"]
e= request.form[ "Fat"]     f=
request.form["Sodium"]     g=
request.form[ "Fiber"]     h=
request.form["Carbo"]     i=
request.form["Sugars"]     j=
request.form["Potass"]     k=
request.form[ "Vitamins"]     l=
request.form[ "Shelf"]     m=
request.form["weight"]     n=
request.form["Cups"]

    t=[[int (a1), int(a2), int(a3), int(a4), int(a5), int(a6), int (a7), int (b), int(c), int(d), int(e), int(f)
,int(g), int(h),int(i),int(j),int(k),int(l),int(m),int(n)]]     y = model.predict(t)     return
render_template("prediction.html", z = y[0][0])


if _name_ == "_main_":
    app.run(host="0.0.0.0",port=5000)
```

Final Project Report Template

Github and Project Demo Link:

https://github.com/sriharshitha-08/Cereal-Analysis-Based-On-Ratings-By-Using-Machine-Learning-Techniques

https://github.com/THOKALA-SRAVAN

https://github.com/pradeepa-05/Cereal-Analysis-Based-On-Rating-By-Using-Machine-Learning-Techniques

https://github.com/SaiSravanthi-12?tab=repositories

Video Demo Link:

https://www.youtube.com/embed/uOZtRka6cpE