Data Science Final Project:
Canadian Cheese Fat Analysis by Machine Learning

TRINH DINH LAM
Fall 2024, INFO-6145, Fanshawe College

1. Abstract

Canadian cheeses have always been known for its variety in choices and flavours. There are some typical cheeses that only fit for specific purposes like cooking or baking and not for direct consumption because of fat level in them. Due to its varieties and usages, it is really crucial to correctly determine fat content of the cheeses to address their quality and type. By successfully classifying fat level, manufacturers are able to label their products precisely and follow the industry standards. Additionally, this may allow customers to make better decision for their choice of cheese as well as increase customer's satisfaction for manufacturers' authorized dealers.

2. Introduction

Cheese has been one of the greatest artisan product of humankind and the very first evidence of cheesemaking has been found on an Egyptian art walls last more than 4000 years ago. Throughout centuries, it becomes a very wide-known product to be traded and adopted by all various countries like Europeans, Middle Eastes. With countless varieties, flavours and textures, cheese keep being one of the most important part of our food cultures nowadays.

Despite rich history of being an artisan product made by many complicated techniques, a rise of cheese industrialization has been recorded from the very beginning of 18 centuries making it become massively produced and available everywhere on earth. This trend attracted millions of business and became an increadible industry worth billions of dollar.

Canada joined the cheese industry from the very beginning since the country is famous for its high-quality and exceptional dairy products.

An industry has been observed with gradual growths in recently years with massive amounts of production and consumption along all provinces. A stable rise in cheese Canadian cheese market enables manufacturers apply automated systems in production and this is the factor that make fat level of cheese uncontrolable. The fat is not only a crucial characteristic of cheese product which address its textures and tastes but also an element allow customers to decide whether this typical type of cheese could be served for their purposes. Fat content in cheese could be determined by multiple methods including two of the most popular are Monjonnier and Gerber techniques. Even there are some effective ways to identify richness of cheese, they are required a decent cost of time to be processed. That is the reason why major factories are building Artificial Intelligence technologies in order to make precise prediction of fat level based on numberous inputs and features in manufacturing procedures. Successfully intergrating predictive models in cheese production can be a powerful tool to delegate business save a lot of budget on pointing out fat content of their products based on statistical analysis instead of outdated technique which is costly in term of money and time. Morever, precisely labeling products satisfy business partners, increase their reputation and customers' satisfaction.

The main objective of this study is to build a predictive models to classify fat level of various Canadian cheeses by using several machine learning methods in Google Colab Platform. Based on given characteristics and related attributes like manufacturing methods, milk type of ingredients, milk treatments process,… this research focuses more on the accuracy of predictive models and methodologies in predicting fat content. In the context of Supervised Learning, to reach the goal of this problem, we will try to apply Classification to categorize fat level into discrete data group instead of using Regression to predict continuous values to reach our target in this project.

3. Methodology and Model Deployment

This research conduct fat level of cheese prediction using a dataset "canadian-cheese-directory" which emphasizes numberous kinds of cheese and their specific properties in all the provinces of Canada. This dataset insist of excessive amount of informations of Canadian cheese types and its revelant attributes which can be collected from Kaggle, an open-source web server for Data Science studying and analyzing purposes.

The prediction will employ 6 Machine Learning Algorithms which are Logistic Regression, Random Forest Classification, Support Vector Classification, Neutral Network(MLPClassification), Gradient Boosting Classification, Naïve Bayes Classification. Feature attributes for this prediction will be manufacturing location, moistures percent, organic type, milk type,… in which fat level would be the targeted output. The prediction model will be implemented inside Google Colab using Python programming language. Google Colab is chosen due to its convinience, computing resources and physical resources like GPU and TPU. Before testing and training the model, we will analyse the dataset and handling missing values to reduce the complexity of the model training process. Finally, each model will be compared to others using performance metrics on testing set by accuracy, recall, precision, f1_score, specificity. Consequently, the prediction framework will consist three parts, which are:

- Preprocessing Dataset
- Feature Engineering
- Algorithms

3.1 Data Processing

*Importing Necessary Libraries

As the dataset will be analyzed and predict using Python programming language, We will try to use pandas, numpy, mathplot and sklearn in the study. Pandas allow us to capture raw data from file like csv, excel to a dataframe to be easily worked with. We will use numpy to deal with metrix as well as sklearn to build our predictive models. Finally, mathplot will help us to plot our data for evaluation and comparision

*Loading Dataset to Dataframe

In order to work with dataset, the dataset must be visualized in an organized way with the help of pandas Dataframe module. We can apply read_csv() function to simply read whole csv file to a dataframe. Additionally, pandas can help us get some basic information of the dataset with many statistical values like mean, standard, frequency,… with info() and describe() functions

3.2 Feature Engineering

*Handling Missing Values

The primary process in machine learning consist of data preprocessing and missing data handling for data preparation. The data will be pre-processed by using mechanics of imputation of missing data supported by pandas library on DataFrame. The module of pandas support various types of functions for cleaning missing values even including replacing missing field with mean, or any values that we want.

*Encoding Categorical Variables

In Classification, it is crucial to convert categorical variables to binary format for prediction. In this research, I will use get_dummies() function of pandas library to easily encode all categorical variables in the dataframe.

*Dropping Unused Features

In order to reduce the complexity of training process, it is recommend to drop unnecessary and complex features like CheeseName, CharacteristicEn, FlavourEn and CheeseID in this research.

*Defining X-features and y-target axises

After processing dataset, defining features X-axis and target y-axis would be the next to make predictive models understand the outcome and inputs. Our target in this context would be fat level and we will define remaining attributes as features for prediction

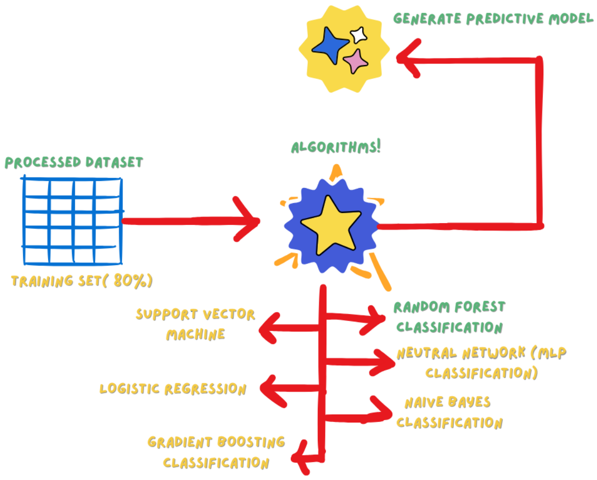
*Distribute Data into Training and Testing

In this study, it is ideal to split the dataset 80% distribution to training and remaining 20% for testing

3.3 Algorithms

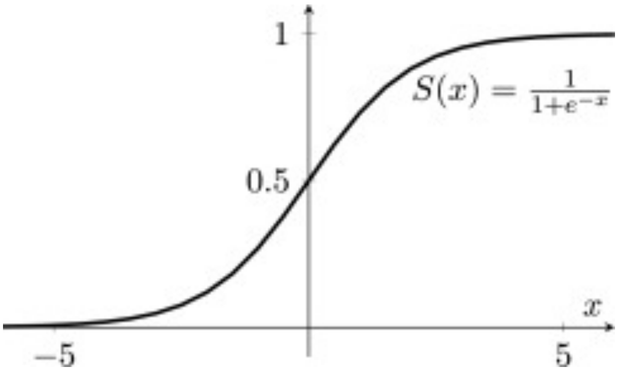
The predictive used a supervised machine learning methodology to predict the fat level. After data was pre-processed, it would be inputted to the learning algorithms. Various features combination was fed into the algorithm to be a candidated feature for the predictive model. And before using the data, dataset were splitted into 2 part: distribution of 80% for training and 20% for testing

The predictive modeling for this study used in this study were Classification methods to predict discrete data instead of regressive predictive models except for Logistic Regression due to its structure for binary classifcation problem and probability outcome provision.

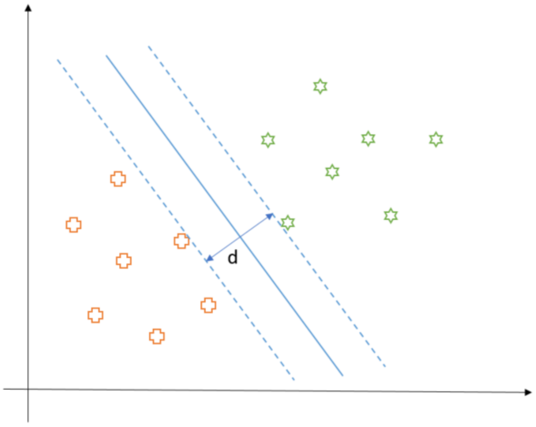

GENERATE PREDICTIVE MODEL
PROCESSED DATASET
ALGORITHMS!
TRAINING SET(80%)
SUPPORT VECTOR MACHINE
RANDOM FOREST CLASSIFICATION
NEUTRAL NETWORK (MLP CLASSIFICATION)
LOGISTIC REGRESSION
NAIVE BAYES CLASSIFICATION
GRADIENT BOOSTING CLASSIFICATION

When it comes to a Classification problem, we can use many kind of algorithms to build predictive model and in this project, I'll try to use 6 methods for our model:

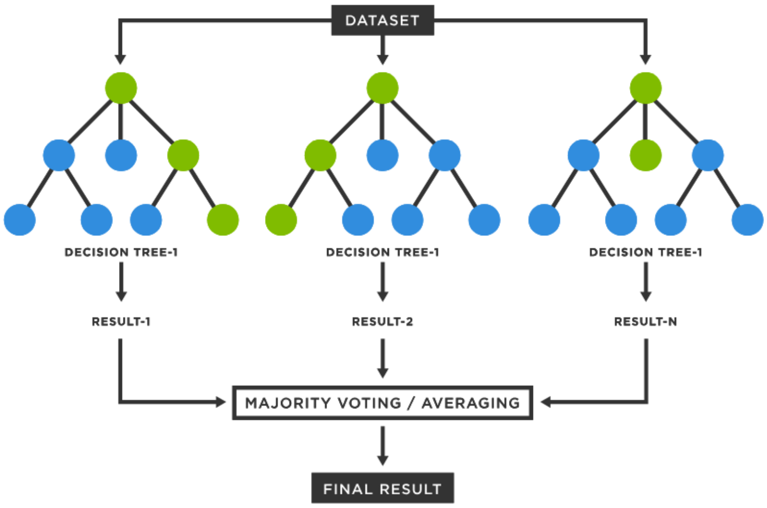
- **Logistic Regression**: basic linear model with high model interpretability, computational effiency and can handle multiple features types. By using sigmoid function, the model can solve Classification problem when dependent variables are discrete or simply binary categorical(yes/ no)

- 

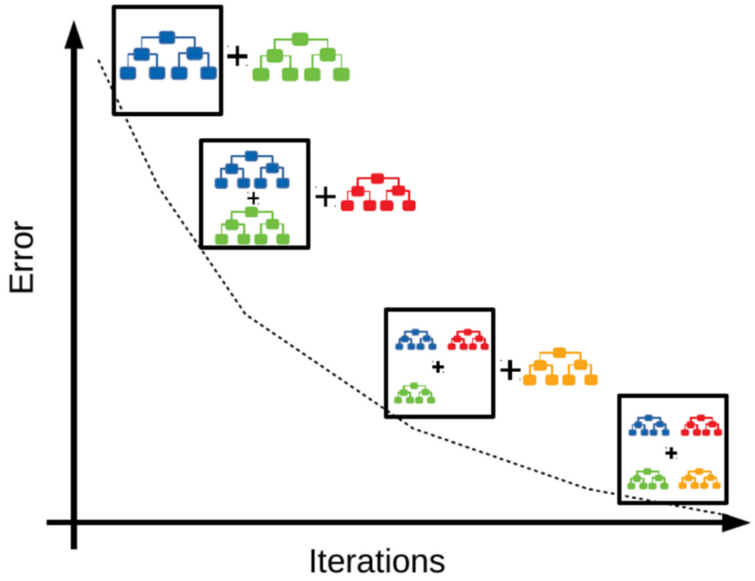
- **Support Vector Machine**: as we have to deal with high dimensional features and maybe non-linear relationships in data. Setting boundaries to classify the data, SVM can be powerful in building model for this problem.



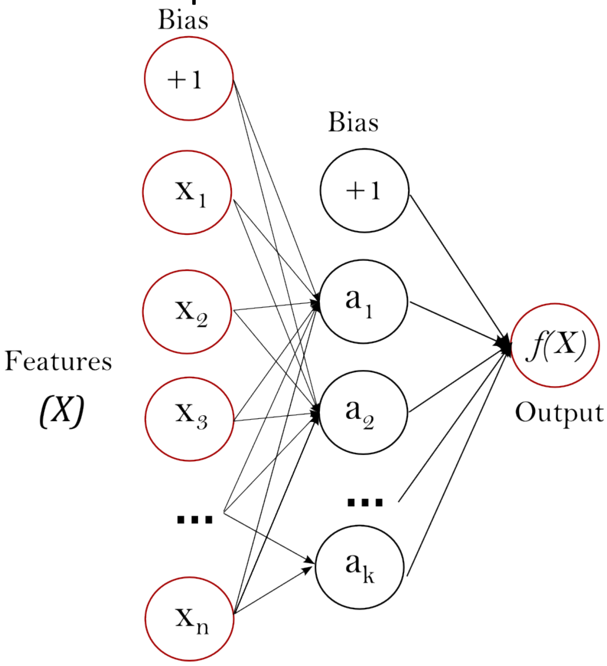
- **Random Forest Classification**: also built from multiple decision trees but adding randomness while splitting the tree help random forest improve better accuracy and effiency in classification problem.



- **Gradient Boosting Classification**: being known with the ability to handle complex relationships in data and against overfitting based on multiple decision tree. The gradient boosting model combines multiple wea learners to predict data. The loss caused by weak learners will be used to calculate the gradient enable model to find the direct of its hyperparameter adjustment to reduce loss in our predictive result.



- **Neutral Network(MLP Classification)**: an excellent model for predict target with complex combination and non-linear relationship of features.



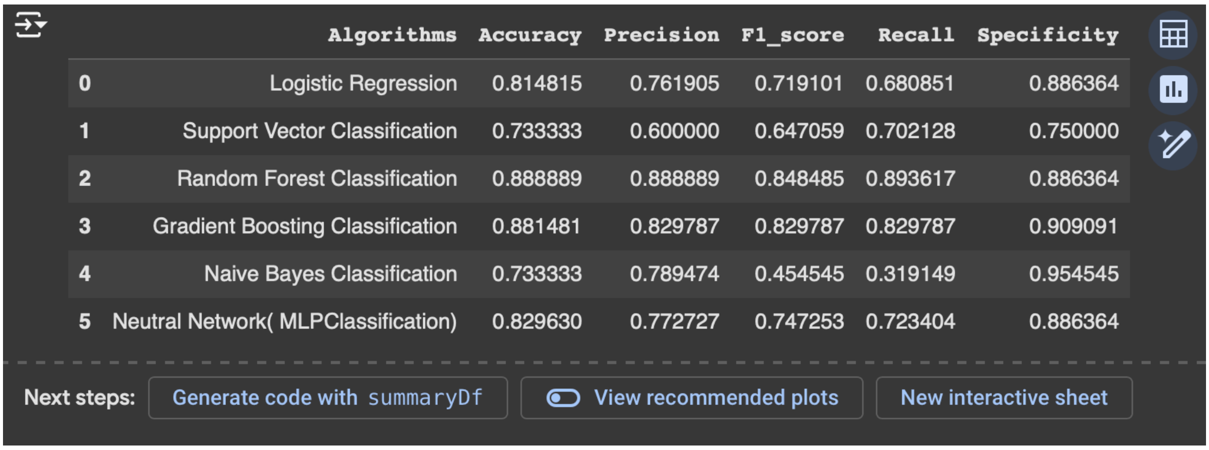
- **Naïve Bayes Classification**: trustworthy model which has highly scalable ability with number of predictors and data points based on probabilistic mathematic.

$$\underset{\text{Posterior Probability}}{P(c\,|\,x)} = \frac{\overset{\text{Likelihood}}{P(x\,|\,c)}\,\overset{\text{Class Prior Probability}}{P(c)}}{\underset{\text{Predictor Prior Probability}}{P(x)}}$$

$$P(c\,|\,X) = P(x_1\,|\,c)\times P(x_2\,|\,c)\times\cdots\times P(x_n\,|\,c)\times P(c)$$

4. Conclusion and Results

- Evaluation Metrics:



	Algorithms	Accuracy	Precision	F1_score	Recall	Specificity
0	Logistic Regression	0.814815	0.761905	0.719101	0.680851	0.886364
1	Support Vector Classification	0.733333	0.600000	0.647059	0.702128	0.750000
2	Random Forest Classification	0.888889	0.888889	0.848485	0.893617	0.886364
3	Gradient Boosting Classification	0.881481	0.829787	0.829787	0.829787	0.909091
4	Naive Bayes Classification	0.733333	0.789474	0.454545	0.319149	0.954545
5	Neutral Network(MLPClassification)	0.829630	0.772727	0.747253	0.723404	0.886364

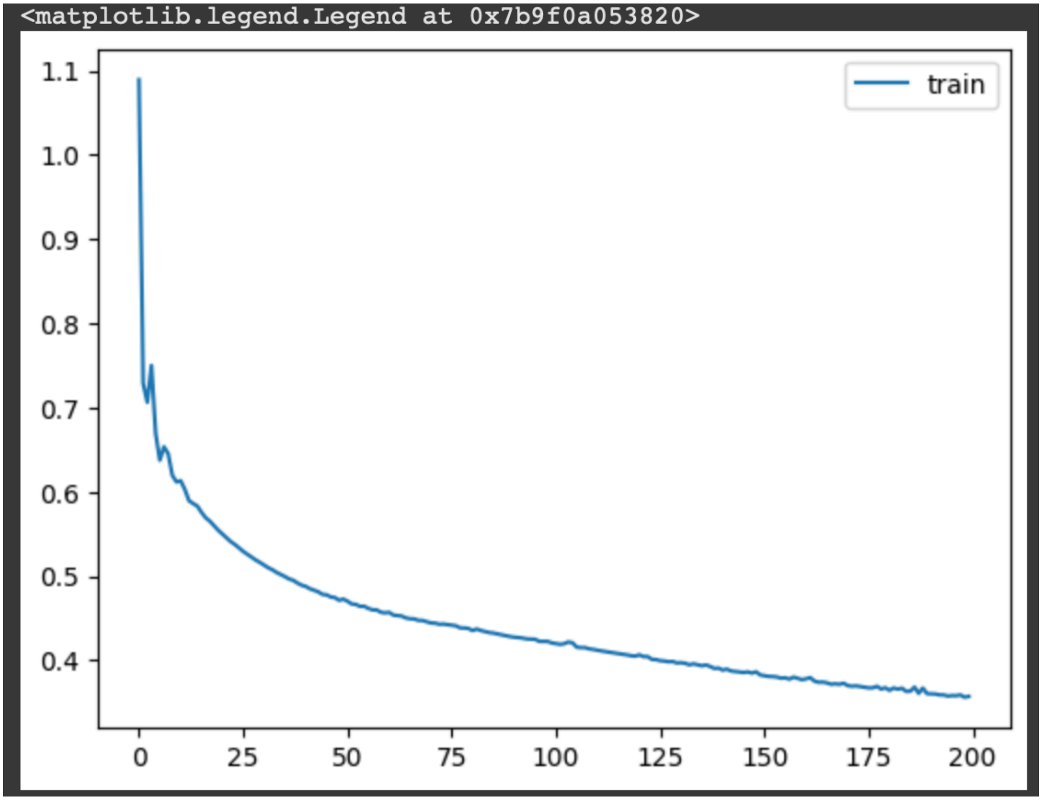
Evaluation metrics of Algorithms

- Conclusion:

The metrics indicates how effective applied algorithms worked on predict fat level in dataset. Overall, Random Forest vs Gradient Boosting are two of the top methods to forecast the target which can be showed in the value of accuracy and precision. The reason why Random Forest and Gradient Boosting could make so much different is they are the combination of many weak learners or decision trees. One decision tree may not give us the best result but an ensemble of trees may give us a very trustworthy vote. Generally, Gradient Boosting is also a good model in comparision with Random Forest.

Hence, given dataset doesn’t seem to be ideal to give Gradient Boosting model best result yet. In the recent study, Random Forest could be the model deployed to be used in predicting fat content of cheese due to its stability and powerful ensemble learning technique. Each decision tree pick random datas from the given dataset to learn from may be underfitting but the result of model is the voting of entire forest which make the model has low bias and virance. This is the reason why Random Forest gave us the best result on this study.

Other than that, 4 remaining models are inferior which can be seen in lower F1_score and recall except for Neutral Network Multiple Layers Perception seemed to give us very reliable informations overtime. In the figure below which captured in the training process of MLP model, we can actually see the loss function decrease when MLP Classifier try to utilize the number of iterations inside model.



```
<matplotlib.legend.Legend at 0x7b9f0a053820>
```

Finally, to improve predictive outcomes in the future, we can keep continuing train MLP and Gradient Boosting when our dataset become vast because their architecture were made to deal with very complex and intricate data.

5. Preferences

- https://www.kaggle.com
- https://ortoalresa.com/en/determination-of-fat-content-in-milk-and-milk-products-for-quality-control/#
- https://www.statista.com/outlook/cmo/food/dairy-products-eggs/cheese/canada
- https://nationalhistoriccheesemakingcenter.org/history-of-cheese/