COMPARISON OF MACHINE LEARNING ALGORITHMS FOR MEAT SUBSTITUTE COMPANIES IN MARKET EXPANSION

LUQI LIANG

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

WORD COUNT: 8043

STUDENT NUMBER

2060460

COMMITTEE

dr. Peter Hendrix
prof. dr. Çiçek Güven

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 23, 2022

ACKNOWLEDGMENTS

I sincerely thank dr. Peter Hendrix for his guidance and help throughout this research process, which would not have made many breakthroughs without him. I would also like to thank dr. Çiçek Güven in advance for being the second reader of this thesis.

COMPARISON OF MACHINE LEARNING ALGORITHMS FOR MEAT SUBSTITUTE COMPANIES IN MARKET EXPANSION

LUQI LIANG

Abstract

There is currently a growing demand in the plant-based meat industry. One of the business strategies for a plant-based meat company is to gain market share by exploring a new geographical location. Existing studies have proven that product features play an important role in predicting as well as stimulating consumers' buying intentions. While the implementation of machine learning algorithms is often overlooked in this domain, the present research aims to classify the impact on geographical decision-making through a comparison between various machine learning models. The main research question is *"What is the impact of product features on the geographical market decision of plant-based meat companies when entering a new market?"*. The present research distinguishes itself by making the use of machine learning algorithms with the combination of all product features in the plant-based meat domain. The models include Naïve Bayes (NB), K-Nearest Neighbors (KNN), Random Forest Classifier (RFC), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGB). The dataset is collected from Mintel, which carries characteristics of plant-based meat products that are sold in different countries worldwide. As a result, NB stands out with a balanced accuracy score of 0.48 and a weighted F_1 score of 0.61 to become the most ideal model for implementing this study. In addition, the product features of "company", "brand" and "claims" take over the top three features that hold the strongest performance in terms of predicting the market location.

1 DATA SOURCE/CODE/ETHICS STATEMENT

Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data and code used in this thesis was obtained from an external organization, which retains ownership of the data and code during and after the completion of this thesis. The author of this thesis acknowledges that they do not have any legal claim to this data or code. The code used in this thesis is publicly available with subscriptions required on https://www.mintel.com/.

2 INTRODUCTION

The goal of this research project is to discover if there is a relationship between product features and the geographical locations that plant-based companies plan to enter. To achieve this goal, various machine learning algorithm models are examined and compared for optimal results.

With the rapid growth in the plant-based meat industry, companies are expanding into different markets worldwide with specific plant-based meat products. For example, plant-based chicken, plant-based burgers, plant-based salmon, etc. Therefore, the market expansion strategy for each plant-based meat company in each country draws a great amount of attention from both meat substitute companies and regular meat producing companies. When a plant-based meat company enters or introduces a food category into a new market, product features can be considered a strong indicator of success. Finding the relationship between those features and the geographical market a company has entered therefore becomes crucial, however, not impossible. Lazzarini, Visschers, and Siegrist (2017) discovered that country of origin has an influence on consumers' perceptions of plant-based foods in terms of sustainability, but the idea of comparing all the product features with the product's selling country is a relatively new topic.

The societal relevance of this project is that by applying machine learning algorithms to existing product features and the geographic market location a company chooses to enter, both meat substitute companies and regular meat-producing companies will gain a better vision. It will not only be beneficial for their own market decision-making process, but also for their competitor's strategic point of view. As previously stated, current literature focuses on product features that could be an indicator of consumers' buying power for both meat products and meat substitutes in different countries. And companies should use different marketing strategies in different geographical market locations (S. M. Kim & Park, 2020). The academic contribution of this thesis is that it addresses and predicts the relationship between product features and geographical market locations for plant-based meat companies by using machine learning algorithms, which is significantly different from the existing predicting methods.

The main research question that this research will address is:

> *What is the impact of product features on the geographical market decision of plant-based meat companies when entering a new market?*

To answer the main research question, the answers of the following two sub-questions will contribute:

SQ1 *Which machine learning algorithm is a good fit for discovering the relationship between product features?*

To discover the relationship between features, modeling through five machine learning algorithms will be carried out. They are Naïve Bayes, K-nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGB). Because all features can be categorized into discrete class labels, which indicates a typical classification problem, the baseline of this search will be set by running the Naive Bayes algorithm.

SQ2 *Which product features have the most impact on the decision of entering a market?*

A set of features that are partly based on previous research will be served as predictors and the feature of market will be the target variable. Feature ablation will define the features that deliver the best prediction in terms of finding the relationship with the target variable.

The findings of this study confirm that NB exceeds other four algorithms to deliver the best prediction performance with the given product features. It is mainly because the nature of this algorithm includes dealing with imbalanced dataset and tendency to assume a strong feature independence. After model evaluation, feature performance therefore proceeds on the NB model. As a result, product features demonstrate the highest predicting performance after having removed the features of "company", "brand" and "claims".

3 RELATED WORK

As the aim of this study is to estimate predictions on product features and the geographical market that plant-based meat companies plan to enter, related work mostly contributes on the importance of various product features and machine learning approaches.

3.1 *Product Features*

Kerslake, Kemper, and Conroy (2022) suggested that product packaging, labelling, for example, carbon footprint logos and certifications and an image of the product on the packaging of the meat substitute product can increase buying intentions and stimulate consumers' willingness to buy. On the other hand, high product price and bad brand reputation also show a negative effect on the buying process (Carlsson, Kataria, & Lampi, 2022). While sensory aspects might not be relevant for consumers, versatility of the product is found to be another facilitator. Another research conducted by Alanís et al. (2022) revealed that even for meat product like sheep meat, consumers in Mexico perceive it as unique sensory attributes and are willing to pay for a higher product price when there are certain claims on the package such as certified organic. Another attribute influences Mexican consumers in buying sheep meat is their educational level. Targeting Mexican consumers in Canada or the United States brings meat selling companies more potential. Nevertheless, consumers in different geographical locations seem to have different appetites for plant-based meat products. For example, in Sweden and other countries, sensory quality and added value of non-meat products appear to be the barriers in choosing meat substitutes over meat (Collier et al., 2021). For consumers in New Zealand, the more convenient a meat substitute is for them to prepare, the higher chance they would buy. While in Germany, product price plays a key role in local consumers' decision-making process regarding plant-based products (Lemken, Spiller, & Schulze-Ehlers, 2019). Foreign meat companies are also recommended to target consumers' sensory experience, purchasing locations as well as seasonality when launching products to the Chinese market (Kantono, Hamid, Ma, Chadha, & Oey, 2021). It is rather obvious that different marketing strategies (S. M. Kim & Park, 2020) and market driven initiatives (Sandøe et al., 2022) should be established for the global market. The work of this research aims to evaluate market driven initiatives, such as product features, by running machine learning algorithms to classify the impact on the geographical decision making that plant-based meat companies chose to export to.

3.2 *Machine Learning Approaches*

H. He, Sun, Li, and Mensah (2022) proposed a comparison between different prediction models including Logistic Regression, Multilayer Perceptron (MLP), Support Vector Machine (SVM) classifier in terms of predicting effect of crude oil prices. The study later found out SVM is the most effective way in predicating such a scenario, which is also identified as a classification problem. In another research that was conducted by Y. Chen and Zhang (2022), Support Vector Machine (SVM) and Multinomial Naïve Bayes (MNB) are stated as the state of art in terms of forecasting the Chinese consumers' perceptions on meat substitutes. After carrying out a comparison between eight different machine learning algorithms, van den Bulk et al. (2022) concluded that Support Vector Machine (SVM) and Naïve Bayes (NB) seem to be the models that deliver the best performance in systematic review on food safety, which is also defined as supervised classification problem. K-nearest neighbor (KNN) algorithm has been a popular choice for classification predictions according to Xing and Bei (2020). The study also investigated that KNN is able to deliver even better performance score when it is class-weighted. However, this result is extracted from medical health data which belongs to single-class classification while the problem that this research paper deals with is a multi-class classification. Another research was suggested by Deng, Zhu, Cheng, Zong, and Zhang (2016) that the KNN algorithm is not only a highly efficient learning algorithm for large scale dataset, but delivers even better accuracy on classification predictions. Random Forest (RF) algorithm has proven to be a non-parametric statistical estimation that is widely chosen in machine learning strategies (Athey, Tibshirani, & Wager, 2019). The RF classification tends to improve accuracy by reducing the number of data collected that is caused by its burden (Speiser, Miller, Tooze, & Ip, 2019), and its attributes are carried out based on the high learning activities with low demands in hyper-parameter tuning (Gomes et al., 2017). Carranza, Nolet, Pezij, and van der Ploeg (2021) also believed that the RF algorithm yields slightly better performance score than process-based models in terms of predicating root zone soil moisture, but lacks the capacity of estimating extreme moisture conditions. As a classic ensemble learning method, Extreme Gradient Boosting (XGB) results in the highest evaluation score compared to other methods, such as SVM, RF and KNN, for establishing a classification model for medical purposes (S. He, Li, Peng, Xin, & Zhang, 2021). Other related work by H. Li, Cao, Li, Zhao, and Sun (2020) indicated XGB algorithm is superior in selecting features and achieving accuracy when competing with various tree-based algorithms, as well as Logistic Regression in classification predictions. M. Kim et al. (2022) embraced a similar classification approach on clinical data which is imbalanced. They compared the evaluation results between SVM, RF and XGB, in which XGB produces not only better discrimination, but also more advanced robustness among other chosen ensemble machine learning models. This study outcome also pointed out that ensemble models like XGB do have an effect in predicting cardiology for patients. However, the gap between using machine learning techniques and the specific prediction in terms of market expansion is still vacant in existing literature. This is what makes this research stand out – by combining machine learning models for classification problem and its prediction on geographical market expansion in plant-based meat domain.

3.3 SMOTE

The nature of imbalanced data reveals that various data points are not equally distributed in a certain scenario, which typically leads to the under-representation of one or multiple classes. The minority class often indicates insights of the data source where produces the learning difficulties. It is crucial to find the balance for the varying relationships between different data points that contain multi-class data. Synthetic Minority Oversampling Technique (SMOTE) has shown to be one of the methods for improving the imbalanced distribution for imbalanced datasets (Kerslake et al., 2022). Another study conducted by Hussein, Li, Yohannese, and Bashir (2019) argued that SMOTE is a critical modification for datasets that are highly imbalanced. The synthetic samples are led closer to the data points that are the minority rather than the data points that are the majority. Therefore, the synthetic samples are located within the class distribution. The SMOTE algorithm rebalances the original dataset by applying various replications of classes that belong to the minority, thus creating new synthetic examples. This procedure concentrates more on the value and the relationship between features rather than the whole data points. According to Fernández, García, Herrera, and Chawla (2018), SMOTE works by first calculating the number of oversampling, which can be adjusted by a wrapper process or as one versus one proportion of class distribution. Then, the algorithm randomly determines a minority class, followed by gathering the minority class's K-nearest neighbors. In the last step, an interpolation is carried out. Nevertheless, SMOTE is also famous for being attributable to adding noise and prolong training time (S. He et al., 2021).

4 METHODOLOGY

In this chapter, five selected machine learning models are discussed, as well as the justifications and comparisons among them. The first section contains motivations and explanations between different methods that include Naïve Bayes (NB), K-Nearest Neighbors (KNN), Random Forest Classifier (RFC), Support Vector Machine (SVM) and Extreme Gradient Boosting (XGB). The following section presents a brief description of evaluation metrics that are selected for comparison. Figure 1 illustrates the data science pipeline that this research follows through.

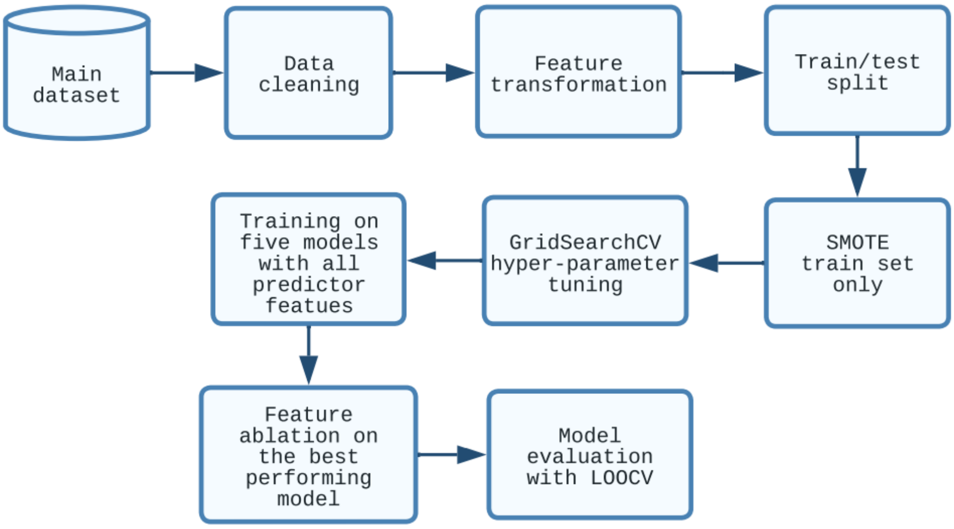

Main dataset
Data cleaning
Feature transformation
Train/test split
SMOTE train set only
GridSearchCV hyper-parameter tuning
Training on five models with all predictor featues
Feature ablation on the best performing model
Model evaluation with LOOCV

Figure 1: General workflow of current project

4.1 *Models*

4.1.1 *Naïve Bayes*

Naïve Bayes (NB) is a classification technique that is famous for being easy to build and simple to run, particularly for large datasets. The way this algorithm works is established on Bayes' Theorem which assumes every predictor is to be independent from each other. NB is often chosen to run when the input dimension is high, and performs well in multi-class predictions (Meenakshi & Geetika, 2014). As a probability-based method, NB reveals a tendency to perform poorly with decision tree method. However, the model itself only needs a small amount of training during parameter estimation, which is another advantage for dealing with classification problems. One of the disadvantages of NB includes assigning zero probability to categories that are not foresaw in training data, which can possibly lead to a limitation in predictions. Another downside of using NB is its assumption of predictor independence which is almost always impossible in real-life scenarios.

Despite the weakness of NB, the value that this model can bring has made it the ideal model for setting the baseline for this study. The main reason is that NB is less time-consuming and easy to train on and sometimes can surprisingly outperform other machine learning algorithms that are with great complexity. To be more specific, the NB model that is chosen for this study is Gaussian Naïve Bayes (GNB). Ampomah et al. (2021) argued that just like NB, GNB also works as a probabilistic classifier and possesses the same simplicity and time-efficiency as the NB algorithm with strong feature independence. The only difference is that GNB is made for classification problems that carry features that are continuous and follows a normal distribution. X represents each observing variable while y consists of all the classes. At each data point, P(x_i | y) symbolises the probability if x is from class y distribution. The distance between that probability to each class mean is summed up by z-score at each data point, which is formulated as follows:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{1}$$

4.1.2 *K-Nearest Neighbors*

As a supervised algorithm, K-Nearest Neighbors (KNN) is widely implemented in machine learning process (Zhang, Li, Zong, Zhu, & Cheng, 2017). KNN is prone to catch different data that share a certain level of closeness by calculating the distance between different data point on a set. For example, Euclidean distance is one of the most popular measures of distance by counting the length of the straight line. The KNN algorithm is suitable for both classification and regression prediction problems. After being provided with a positive integer k, KNN starts to take actions in finding k observations that are the closest to a certain unseen data point from the test data. Then, the algorithm evaluates the conditional probability that the test observation belongs to, followed by assigning the class to where contains the greatest number of data points among all classes.

Syamsuddin and Barukab (2022) considered that the KNN algorithm is beneficial for processing in classification tasks because of its nature of being simple to implement, less relying on parameter tuning and being flexible with adding new data. The concerning part of KNN constitutes three issues. Firstly, the KNN algorithm performs poorly on large dataset due to the cost of estimating the distance every time. Secondly, KNN finds difficulty in working with high dimensions and often needs feature scaling. Thirdly, it becomes necessary to manually adjust missing data and outliers in case of the sensitivity of KNN (Sun, Du, & Shi, 2018).

4.1.3 *Random Forest Classifier*

The Random Forest (RF) algorithm is a tree-based learning algorithm that is ensemble and averages over several individual trees (Schonlau & Zou, 2020). The Random Forest Classifier (RFC) is made of decision trees that are randomly chosen from the subsets of training data. RFC collects the votes from decision trees in order to determine the most suited class for the observations. To reduce the overfitting problem that decision trees can potentially cause, RFC executes bootstrap or the bagging method, thus resulted in a more accurate estimate (Speiser et al., 2019). Parameter tuning is crucial for RFC as it can produce lower errors that are generally induced during training process. In a research on credit card debt default that is conducted by Schonlau and Zou (2020), RFC is proven to have higher accuracy in such classification prediction than parametric models, such as Logistic Regression. This study also found out a better model performance is also possible when the model is dealing with multi-class datasets.

Being one of the most precise learning techniques for classification problems, RF is able to handle and run seamlessly on large datasets without feature deletion. This algorithm also provides estimation on feature importance while initiating unbiased perception on generalization errors, which makes it to be easily parallelized. When facing a dataset with a large proportion of missing data, the RF algorithm unveils robustness, as well as to outliers and noise (Taşcı, 2019). The downside of this model is being easily biased for categorical features with various quantity of levels (Sattari, Falsafian, Irvem, S, & Qasem, 2020).

4.1.4 *Support Vector Machine*

Another representation of supervised machine learning method is the Support Vector Machine (SVM) algorithm. It is mostly considered in classification problems, although it is applicable to both classification and regression predictions. SVM's working process starts with identifying the total number of variables (n), then plots each variable individually in n-dimensional space with a particular coordinate (Huang et al., 2018). By determining the hyper-planes, SVM divides two classes of data points therefore landing to the highest margin. It is worth to mention that the *kernel* function in the SVM technique is what broadly transforms data into a higher dimensional space and becoming separable. This special feature of SVM has made the algorithm appropriate for both linear and non-linear separation problems.

Leong, Bahadori, Zhang, and Ahmad (2021) pointed out that the SVM method is a popular choice for non-linear problems because the model is robust for missing and noise data, and consequently gives better performance results while being less computationally expensive. Because of the clear separation of margin, SVM shows efficiency in implementing datasets that hold high dimensional spaces. Apart from being relatively simple and flexible to utilize, SVM can also handle both unstructured and structured data and has lower risk of overfitting (Pisner & Schnyer, 2019). On the other hand, the model's training time and complexity tend to be higher for large datasets based on the research from Nalepa and Kawulok (2019). It is pivotal yet not easy to fine-tune SVM's hyper-parameters which includes the *kernel* function, *c* and *gamma*.

4.1.5 *Extreme Gradient Boosting*

An Extreme Gradient Boosting (XGB) machine learning algorithm attempts to generate a gradient boosting framework (T. Chen & Guestrin, 2016). Similar as other models that belong to the decision tree family, XGB is relatively easy to be interpreted and feature-explicit (Wang, Lu, & Li, 2019). XGB consists of two major ensemble techniques - bagging and boosting. Bagging generates replacement for data that are sampled, and operates on them to create aggregation (Breiman, 1996). Afterwards, a plurality vote takes place as the aggregation continuously predicts new classes. Boosting creates new models by iterating and adopting error weights until there is no space left for improvements according to Bentéjac, Csörgő, and Martínez-Muñoz (2021).

Several features have made XGB a favorable ensemble learning method. Firstly, XGB is great at scaling data while showing robustness when handling missing values. Secondly, because of its ability to explicate variable natures, the importance of each variable can be expected (C. Li, Zheng, Yang, & Kuang, 2018). Thirdly, Z. Chen and Fan (2021) discovered that the XGB method gives higher accuracy and is fast to interpret when being compared with the Gradient Boosting model. As reported by Wang et al. (2019), the drawback of XGB contains the possibility to be hard in terms of visualization and sensitivity in parameter-tuning. In addition, considerably large amount of hyper-parameters that can be tuned for this model could also be an obstacle.

4.2 *Evaluation Metrics*

A vital step before determining the right evaluation metrics is to investigate the nature of the dataset. In case of the dataset that is used in this study, it is acknowledged as a classification prediction problem with highly imbalanced data. This is primarily because there is a huge difference with the amount of values between different data points. In this section, each chosen evaluation metric is shortly introduced with the equation that achieves the metric.

4.2.1 *Balanced Accuracy*

Balanced accuracy (equation 2) is effective for both binary and multi-class classification problems. Different from the use of an overall accuracy metric, balanced accuracy often delivers promising results for datasets that contain imbalanced data (Chopra & Dixit, 2021). This metric is done by taking the average of sensitivity and specificity score (Pakravan & Jahed, 2022). The closer the balanced accuracy score is to 1 (or 100%), the more correctly the model is predicting on observations.

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2} \tag{2}$$

Sensitivity (also known as *Recall*) is to estimate the percentage of positive cases that are actually detected out of positive classes by the model (Trevethan, 2017), which is presented in equation 3.

$$Sensitivity = Recall = \frac{TP}{TP + FN} \tag{3}$$

Specificity evaluates the percentage of negative cases that are correctly perceived from the negative classes (Trevethan, 2017). The formula is shown in equation 4.

$$Specificity = \frac{TN}{FP + TN} \tag{4}$$

4.2.2 *Weighted F1 Score*

Both balanced accuracy and F_1 score are advantageous for classification problems represented with imbalanced values (K. Chen et al., 2020). However, F_1 score (equation 5) tends to minimize the difference between recall and precision by focusing more on the positive data points.

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5}$$

Precision is the proportion of true positives that are detected in all positive instances which include both false positives and true positives (Mohammed & Omar, 2018) as indicated in equation 6.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

The weighted F_1 score (equation 7) is computed by averaging the per F_1 score attained from each label and taking each label's actual occurrence into account at the same time. With the weighted mean, F_1 score can classify each label's true instances after being weighted by the amount of values of that given label (J. Chen et al., 2022).

$$Weighted\ F_1 = \frac{1}{k}\sum_{i=1}^{k} \beta_i \cdot F_{1i} \tag{7}$$

5 EXPERIMENTAL SETUP

This chapter can be seen as the bridge of implementing the previously mentioned five models and the final results. Firstly, the dataset that is chosen for this specific topic will be defined. Secondly, the step of pre-processing the raw dataset, including data cleaning and data transformation will be discussed. The following sections will talk about the steps for the actual execution and how the model performance is evaluated. The last part of this chapter will list all the software and hardware that are required for this study.

5.1 *Raw Dataset*

The dataset that is used for this research is the plant-based meat products dataset, which is collected from Mintel. Mintel is a world's leading market intelligence agency that gathers high-quality consumer packaged goods data across 86 markets worldwide.This dataset is extracted in February 2022 and covers about 4,554 observations of plant-based meat product in different markets in the world. 12 product features are listed, including company name, brand, product, sub-category, format type, storage, launch type, claims, package type, package materials, price in euros and price per 100g/ml in euros. The feature of market is the target variable, which makes a total number of 13 features in this dataset. A description of all features involved can be found in Appendix A (page 32). 79 categories belong to the target variable, namely 79 different countries with a minimum of 1 data point and a maximum of 462 data points with median of 18 and mean of 58 (see Figure 2). The distribution of this variable makes it a dataset that consists of imbalanced data, and certain actions will be taken at a later stage.

This dataset is representative and relevant based on three reasons. The first reason is the number of observations this dataset contains – 4,554 which covers the market of meat substitutes worldwide. The second reason is that the data is from the period of February 2019 to February 2022, which represents the most current market insights and makes it valuable to run analysis on. The third reason is that the observations chosen in this dataset are specific and concrete, meaning the sub-categories selected under plant-based are only processed fish, meat, egg products, snacks and meals. The dataset is publicly available with subscriptions required. The file is available in excel format with a total size of 462.6+ KB.

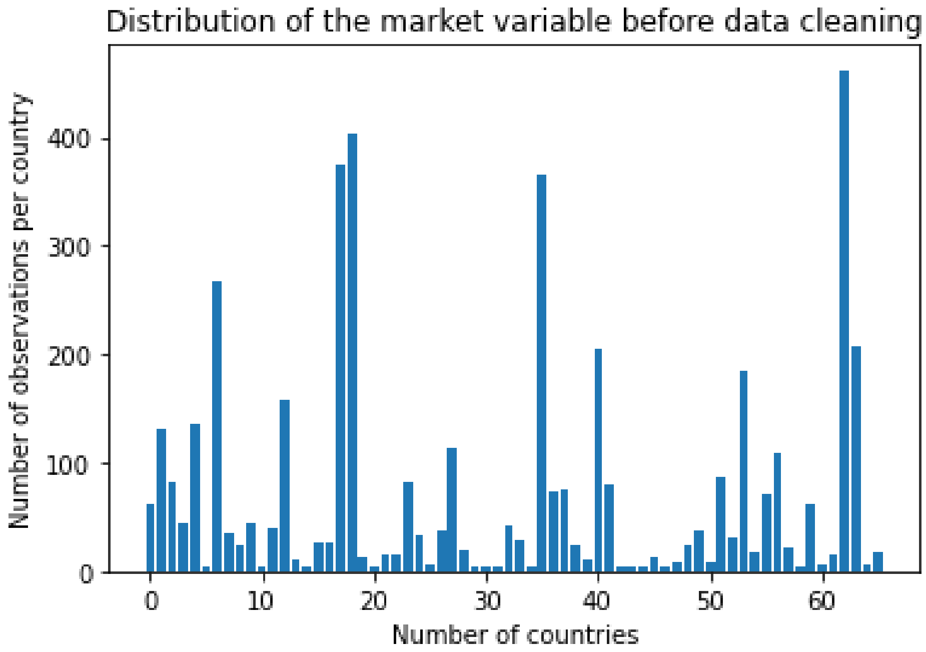


Figure 2: Initial distribution of the market variable

5.2 *Data Pre-processing*

5.2.1 *Missing Values & Outliers*

There are 22 observations with missing values for the "price" feature and 70 observations with missing values for "price_per_100". The missing values from both columns are not a substantial quantity of data compared with the total number of observations, nor are caused by another variable. In order to keep and utilize the most of the data at hand, each missing value is replaced by the mean attained from that column. Statistically speaking, various outliers exist in this dataset. The only feature that is taken into account is the "market" variable, which is also the target variable in this research. This variable demonstrates the country each product has been exported to. There are several outliers displayed for this column. However, only data points that contain one class are deleted for the smoothness of data processing and modeling (XGB) at later stage. All other outliers are kept because they represent different countries and still be in possession of valuable information.

5.2.2 *Oversampling*

As shown in Figure 2, the target feature demonstrates extreme distribution, which holds the possibility to cause bias and negatively influence the performance in data modeling. To prevent this issue, oversampling is carried out by applying the Synthetic Minority Oversampling Technique (SMOTE). This technique works as determining a minority class at random, then gathering the minority class's K-nearest neighbors (Fernández et al., 2018) as mentioned in Section 3.3. The highest value for the target variable in training data appears to be 365. In addition, there are 55 out of 66 data points that own values below 100. Therefore, the oversampling is done manually by increasing the number of values per class that belong to the 55 data points to 100. Based on the fact that the smallest number of the minority class is 2 (after the deletion on 1), the *k_neighbors* parameter of this function is set to be 1. By doing this, the dataset is more equally distributed than before as displayed in Figure 3, with median of 100 and mean of 118 data points per observation. Undersampling is not applicable in this case due to the limited amount of data points for each class.

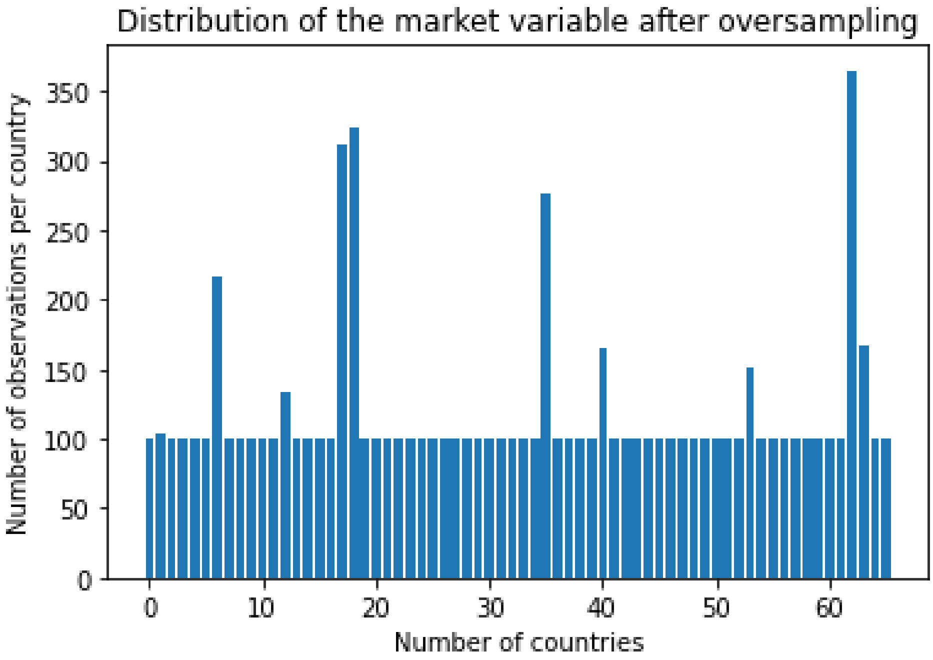


Figure 3: Distribution of the market variable after oversampling

5.2.3 *Feature Transformation*

There are multiple features that are required for feature transformation in order to fit into selected models, for instance, "company", "product", "brand", "sub_category" and more can be found in Appendix A (page 32). The reason behind it is because it is essential to encode categorical data into numeric data before data modeling. Due to the fact that features with high dimensional data can result in a huge amount of categories. Therefore, different encoder methods are considered for two groups of data, namely, *LabelEncoder* for the target variable "market" and *OneHotEncoder* for the remaining categorical variables.

Label Encoding. There are 79 countries included under the target variable "market", which makes it a relatively high dimensional variable. Various markets appear less than 2 times in the whole dataset, which are removed as mentioned in Section 5.2.1 due to sign of being unrepresentative. Label encoder works in a comparably simple and easy manner than other encoding techniques. This encoding technique transformers the labels of data selected from a specific column into a unique integer that is generated from an alphabetical order. By doing this, the "market" variable is converted into numeric values which belongs to a machine-readable form. Therefore, the algorithms that are used in data modeling stage is then able to handle those values in a better way.

One Hot Encoding. One of the most commonly used encoding approaches is one hot encoding (otherwise named dummy encoding). Despite the different names, this encoder strategically creates a new variable (a dummy variable) based on the amount of the unique categorical values. Each value is then assigned judged by its true or false nature with 1 or 0 and added into the column. One hot encoding technique is particularly beneficial for variables that do not follow a intrinsic ordering, nor can be ranked, namely nominal variables. Therefore, all the nominal variables are converted into numerical variables by implementing the one hot encoding method. To be more specific, 1 is appointed to the dummy variable appearing in the dataset and for those variables that do not are left with number 0.

5.3 *Experimental Procedure*

5.3.1 *Train-Test Splitting*

In order to examine the performance of the chosen machine learning models, the whole dataset is divided in to training data and test data. The former is for the purpose of fitting the model while keeping the test data aside. Afterwards, algorithms use the input factors learned from the training data to execute predictions on the test data, followed by evaluations on performance. There are two reasons to omit the k-fold cross-validation before the training process for this study. The first reason is that this dataset is sufficiently large to afford sparing enough data for both sets. The second reason is that this train-test split technique can achieve computational efficiency in comparison with k-fold cross-validation. As a result, the dataset is split into 80% training data and 20% test data with the parameter *random_state* set to be 880. The prediction performance for each model is carried out on the test data, which is left unseen during the training process.

5.3.2 *Feature Ablation*

One of the most important procedures is finding the level of importance for each feature as selecting the predictor variables that reveal the strongest relationship with the target variable is crucial. This procedure can provide an elementary degree of feature interpretability. Tree-based machine learning models possess an instinctive way of discovering feature importance while other models do not, such as the *rbf* kernel for SVM. Feature ablation is chosen for evaluating feature importance for this particular study because of its simplicity, efficiency, and this method is rather practical to resolve for commercial purposes. The application of feature ablation operates as removing one feature from the training data before training the model as the first step. The next step is to estimate the prediction score on the test data. In the last step, the feature importance is calculated by computing the score achieved with all features minus the score achieved without a certain feature. Afterwards, all features are ranked based on the rule that the higher the feature importance, the greater the drop in model performance when the feature is omitted.

5.3.3 *Hyper-parameter Tuning*

Before executing the modeling procedure, tuning hyper-parameter becomes an inevitable step to do in order to enhance the model performance. The technique that is used for finding the optimal hyper-parameters for all the models is called Grid Search, which can be found from the Scikit-learn's library for Python. By giving a dictionary of the values of certain hyper-parameters, predefined hyper-parameters are created. The *GridSearchCV* function then loops through all the combinations of the included values on the training data. The result of this function is to deliver the combination that gained the highest performance score on the predefined method of scoring. As described in Section 4.2, weighted F_1 score is set as the scoring method throughout all the grid search procedure on the five models.

5.4 *Evaluation Metrics*

5.4.1 *Feature Evaluation*

To assess the performance of the classification models on predictive features, two evaluation metrics are applied: balanced accuracy and weighted F_1 score (see Section 4.2).

5.4.2 *Model Evaluation*

To offer the validity on the best operating machine learning algorithm, the Leave-One-Out Cross-Validation (LOOCV) evaluation technique is carried out. LOOCV is well-known for providing robust and precise model evaluation while being computationally expensive. After narrowing down to the most desirable classification model, LOOCV then classifies this model by executing the data that is unseen during the training process in regard to the number of split subsets (k). In the case of LOOCV, the value of k amounts to the quantity of examples in the entire dataset.

5.5 *Software & Hardware*

The software that is used to conduct this research is Jupyter Notebook (6.3.0) from Anaconda, together with Python as the programming language. The proprietary application that is operated to support the coding part is the Intel(R) Core(TM) i5-8265U CPU @ 1.60 GHz @ 1.80 GHz with 8 GB of installed RAM. The list below unfolds the packages and their versions that are contributed.

- *Python (3.8.8)*
- *Anaconda (4.12.0)*
- *SKlearn (1.0.2)*
- *Numpy (1.20.1)*
- *Xgboost (1.6.0)*
- *Category_encoders (2.4.0)*
- *Imblearn (0.9.0)*
- *Matplotlib (3.3.4)*

6 RESULTS

This chapter is devoted to: (1) Deliver the optimal sets of hyper-parameters after tuning on *GridSearchCV*. (2) Provide the results of the baseline model which is the Naïve Bayes (NB) algorithm in comparison with K-Nearest Neighbors (KNN), Random Forest Classifier (RFC), Support Vector Machine (SVM) and Extreme Gradient Boosting (XGB). (3) Present the findings on the best predictor features and weak predictor features, together with their scoring on the chosen model. (4) Give an error analysis for the best performing model by applying with the k-Fold Cross-Validation.

6.1 *Hyper-parameter Tuning*

The hyper-parameter tuning for grid research on NB contains two elements, *var_smoothing* and *cv*. The parameter *var_smoothing* is the proportion of the largest variance of all variables assigned in order to receive the outcome for stability, while *cv* is the number of cross validation that goes through each set of parameters. As shown in Table 1, the range set for the first parameter starts from 0 and ends till -3 with 100 samples to extract from on a log scale. After running through 3 folds of cross validation in linear space, the most optimal hyper-parameter for this model is 2e-06.

Table 1: Parameters of Gaussian NB tuned with GridSearchCV

Algorithm	Parameter	Attempted values
NB	*var_smoothing*	$logspace(0, -3, num = 100)$
	cv	3

There are two parameters assigned for the second selected model KNN, namely *n_neighbors* and *cv* as described in Table 2.

Table 2: Parameters of KNN tuned with GridSearchCV

Algorithm	Parameter	Attempted values
KNN	*n_neighbors*	$range(1, 40)$
	cv	3

The first parameter initiates the grid search technique to work through the number of neighbors from a range of 1 to 40.

Afterwards, the defined number processes on 3-fold cross validation 39 times and delivers 1 as the most suitable number of neighbors for KNN. In order to identify this result, an error rate, together with a rate of accuracy based on the number of neighbors (K) taken from 1 to 40 are shaped sequentially. The minimum error discovered is approximately -0.60132 at when K equals to 0 as shown in Figure 4 with a tendency to increase alongside the value increase of K. The maximum accuracy achieved is presented in Figure 5, which is roughly -0.39868 at K equals to 0.

When the number of K value goes up, the accuracy appears to rise as well. To conclude, the minimum K value that is possible to utilize is 1 according to the findings proceeding from the three techniques discussed above.

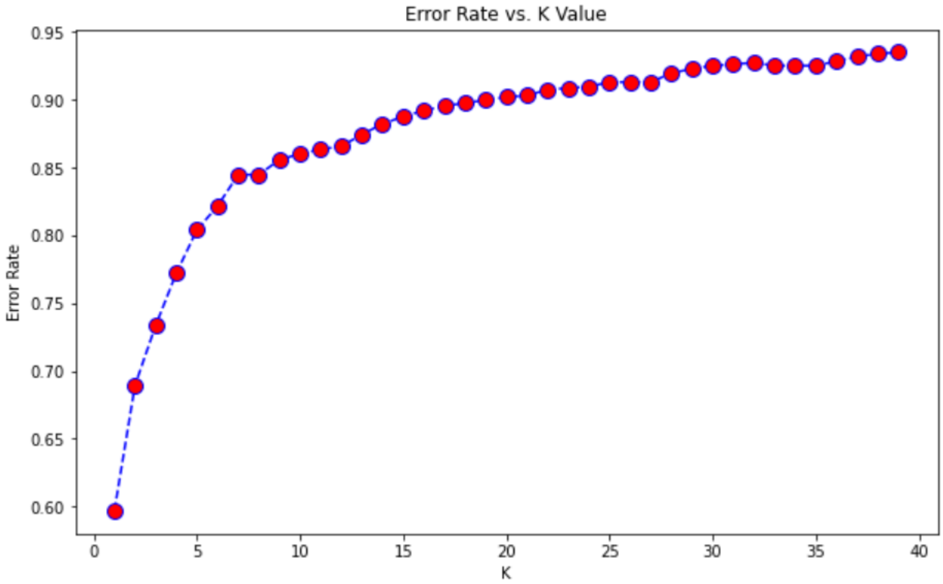


Figure 4: Plot between error rate and optimal K value

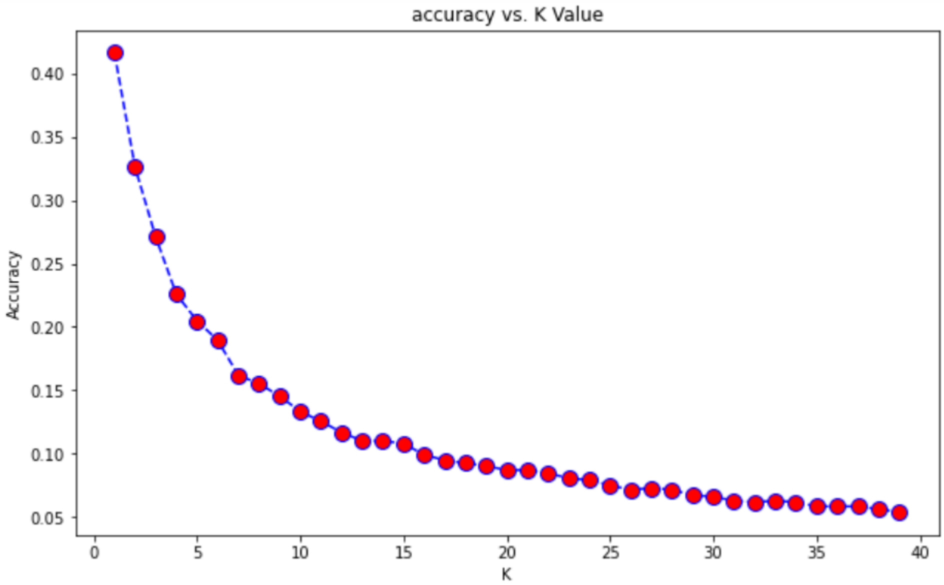


Figure 5: Plot between accuracy and optimal K value

For the RFC algorithm, four parameters are considered as demonstrated in Table 3. They are consisted of the number of classifier estimators (*n_estimators*), the maximum number of features (*max_features*) and depth (*max_depth*) for the tree split, and ended with the function for estimating the split quality (*criterion*). The outcome of carrying out grid search on those parameters includes 100 trees in the forest with maximum 75 depth calculated by the 'log2' function to look for the most desirable split. Last but not least, the combinations for the tree split quality are evaluated by the 'gini' function.

Table 3: Parameters of RFC tuned with GridSearchCV

Algorithm	Parameter	Attempted values
RFC	*n_estimators*	[50, 100]
	max_features	[auto, sqrt, log2]
	max_depth	[35, 50, 75]
	criterion	[gini, entropy]

Kernel, *C* and *gamma* are the three parameters for the grid search on SVM as described in Table 4. They respectively indicate the kernel type, the regularization parameter and the function for calculating the kernel coefficient. The optimal hyper-parameters in this case are 'linear' for *kernel*, 1 for *C* and 0.1 for *gamma*. In other words, linear function is proved to deliver the highest classification accuracy with a penalty error of 1, which then creates an influence of a data point with a distance of 0.1.

Table 4: Parameters of SVM tuned with GridSearchCV

Algorithm	Parameter	Attempted values
SVM	*kernel*	[poly, rbf, sigmoid, linear]
	C	[0.1, 1, 10]
	gamma	[0.1, 1, 10]

One of XGB's drawbacks is showing less sensitivity in terms of parameter-tuning compared with other machine learning algorithms (see Section 4.1.5). Therefore, more parameters are implicated under grid search for this model. The *subsample* parameter is to generate the ratio of training data while *colsample_bytree* is for the ratio of columns for each tree. To produce the maximum number of nodes in terms of complexity and the minimum weight for generating new nodes, *max_depth* and *min_child_weight* are determined. As two essential parameters for the XGB algorithm, *learning_rate* and *n_estimators* verify the speed of the learning process for the model and how many trees the model performs on. A general trade-off between those two parameters is that the lower the learning rate, the more trees are desired to develop the model. After tuning, the hyper-parameter set that stands out is 0.75 for *subsample* and 1 for *colsample_bytree* with a maximum nodes amount of 35 and minimum nodes weight of 5. Moreover, a learning rate of 0.5 in combination with 500 trees to test on seems to be comparably ideal as well.

Table 5: Parameters of XGB tuned with GridSearchCV

Algorithm	Parameter	Attempted values
XGB	*subsample*	[0.3, 0.5, 0.75]
	colsample_bytree	[0.5, 0.75, 1]
	max_depth	[10, 35]
	min_child_weight	[1, 5]
	learning_rate	[0.1, 0.3, 0.5]
	n_estimators	[100, 500]

6.2 *Model Performance*

In this section, the classification performance for the feature types mentioned in Section 4.1 on plant-based meat products is presented. Due to the characteristic of the dataset that is highly imbalanced, Synthetic Minority Oversampling Technique (SMOTE) is applied (see Section 5.2.2) in order to balance out per observation and improve per model performance. Accordingly, the model that is selected to work as the baseline model for this particular study is NB. It is because the NB algorithm is explicitly famous for being simple and efficient to run on large dataset, especially when the input dimension is high. There are 12 predictor features and 1 target feature, which can be found in Appendix A (page 32). The model performance is evaluated based on using the training data consists of all 12 predictor features and hyper-parameter tuning for each model.

Table 6 shows the predicting results of the chosen models and the baseline model after applying all the training data of all 12 predictor variables. The evaluation measures are the balanced accuracy and the weighted F_1 score. As shown in the table, the NB approach notably outperformances other machine learning algorithms that are usually known for giving promising results. The estimates for NB, which are also the baseline score are 0.48 on balanced accuracy and 0.61 on weighted F_1 score with all predictor variables. The second highest score for balanced accuracy belongs to the RFC algorithm with 0.41 while SVM yields the second place for weighted F_1 score with 0.57. Both results are lower than the baseline performance. The approach that generates the worst classification performance is the KNN model which yields a prediction score of 0.37 on balanced accuracy and 0.47 on weighted F_1 score, respectively. XGB presents mediocre results compared with the baseline performance, which are 0.40 on balanced accuracy and 0.56 on weighted F_1 score. The baseline model has surprisingly outperformed all other models, which makes it pass through to the next procedure with feature ablation.

Table 6: Scoring performance on five models with all 12 predictor features

	Model Performance	
	Balanced accuracy	Weighted F_1 score
NB	**0.48**	**0.61**
SVM	0.37	0.57
RFC	0.41	0.56
XGB	0.40	0.56
KNN	0.37	0.47

6.3 *Feature Performance*

The purpose of implementing feature ablation procedure is to discover the feature group that achieves the highest performing score by removing one feature at a time. The baseline that is used to compare per feature performance is the score of NB as mentioned in Section 6.2, which is 0.477 and 0.606, respectively. Two identified evaluation metrics are balanced accuracy and weighted F_1 score as mentioned in Section 4.2. The comparison between features takes both metrics into consideration. However, the ranking of feature importance is based more on the estimation of weighted F_1 score as the goal of this research is to classify each country's true instances. Feature importances for the NB model, as estimated by decreases in balanced accuracy and weighted F_1 score when a feature is omitted from the model can be found in Appendix B (page 32). There are six features outplayed the baseline and the six remaining features that underperformed the baseline. Feature importances for the 3 most important features in the NB model, as estimated through decreases in balanced accuracy and weighted F_1 score when the feature is omitted from the mode as shown in Table 7. The performance of those three features on both balanced accuracy and weighted F_1 score are slightly lower than the baseline. All in all, after ablating the "company" feature, the model performance drops 0.069 on balanced accuracy and 0.07 on weighted F_1. The variables that are ranked the second and third are "brand" and "claims" with a decrease in weighted F_1 score compared to the baseline result.

Table 7: Top 3 feature importance on NB with feature ablation

	Feature Performance	
	Balanced accuracy	Weighted F_1 score
company	**0.069**	**0.07**
brand	0.083	0.068
claims	0.022	0.018

6.4 *Model Evaluation*

The NB algorithm stands out as the best performing model for this study. The evaluation of this model with its tuned hyper-parameters takes place by applying the Leave-One-Out Cross-Validation (LOOCV) technique as described in Section 5.4.2. This approach includes running a sensitivity analysis on a range of k values between 2 and 30, then compare the findings on mean classification weighted F_1 score. The line plot in Figure 6 presents the comparison between the mean score of weighted F_1 score to the LOOCV score. The error bars are the distribution with the minimal and maximum values for each fold. The horizontal bar is the indicator of the idea test condition.

The appliance of this technique shows the LOOCV result is about 0.730 for implementing the NB model, which is slightly lower than when k equals to 8 (0.733) and 16 (0.731). Additionally, most k values fall right below the idea case as can be seen in the Figure 6.

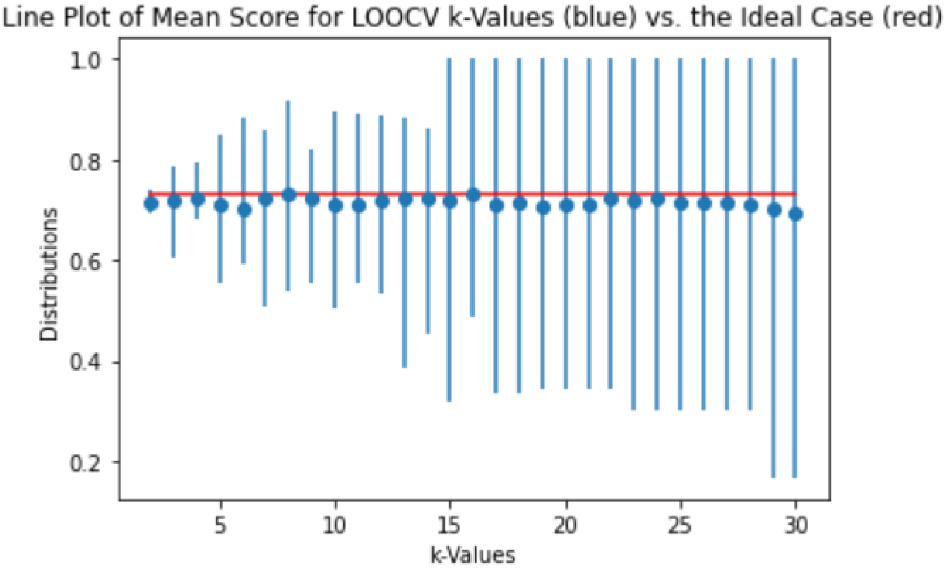


Figure 6: LOOCV line plot of k values and its ideal test condition

This scenario proposes that when performing at most k values, the model tends to underrate in comparison with its performance on the idea situation. With k value amounts to 16, it can be concluded a more optimal evaluation when realizing predictions on new data with weighted F_1 score equals to 0.731.

7 DISCUSSION

The goal of this study is to detect the impact of product features of plant-based meat on the expansion of a company to a new geographical location. The following sections explain the composed outcome, together with the contributions within the existing framework. Furthermore, limitations are discussed in the corresponding section, followed by the recommendations for future work. To sufficiently reveal the findings, the main research question is investigated: *"What is the impact of product features on the geographical market decision of plant-based meat companies when entering a new market?"*.

7.1 *Research Question One*

SQ1 *Which machine learning algorithm is a good fit for discovering the relationship between product features?*

To answer the first sub-question, five machine learning algorithms are chosen based on the existing literature (Section 4.1). For the purpose of achieving more accurate estimation on selected models, hyper-parameter tuning is carried out through grid search. After predicting all the predictive variables on each algorithm, the NB model emerges as the final winner with balanced accuracy score of 0.48 and weighted F_1 score of 0.61, respectively (see Table 6). Balanced accuracy is calculated by taking the average of sensitivity and specificity. With a score of 0.48, it means NB produces an accuracy of 48% to identify both the negative and positive class on average. However, if the aim is more focused on predicting the positives, then the metrics of weighted F_1 score should come to play. With a result of 0.61, it proves that the NB model holds the ability to predict each true label of 61% accuracy. This result is reached by tuning the *var_smoothing* parameter to 2e-06 while enforcing 3 folds of cross validation.

NB as the named baseline notably outperformed more sophisticated models, such as KNN, RFC, SVM and XGB. One of the reasons that it occurs might be the high dimension of the target variable "market". Another reason is the nature of the NB model, which holds the tendency to assume a strong feature independence as mentioned in reviewing the existing literature in Section 4.1.1. It is noteworthy to mention that when the number of neighbours being 1 is demonstrated to be the optimal result for KNN's hyper-parameter tuning (see Section 6.1). This implies overfitting and that the KNN model is estimating on a single sample group only. The reason might be caused by the downside of enforcing this model, which is possessing difficulty in working with high dimensional data as reported in Section 4.1.2.

Limitations of the selection on different machine learning models lie in the severe class imbalance and the curse of dimensionality. In the given dataset, there are 79 data points with observations per class starting from the minimal amount of 1 to the maximum amount of 462. During the training process, as the number of features increases, the sparser the space between different features increases as well.

7.2 *Research Question Two*

SQ2 *Which product features have the most impact on the decision of entering a market?*

The second sub-question is resolved by the execution of the feature ablation procedure. By doing so, a comparison of performance score between different predictive features as well as the baseline performance is gained (Section 6.3). The baseline performance on the NB model suggests 0.4777 for detecting both positives and negatives as well as 0.606 for only predicting the true labels. As previously described in Table 7, the "company", "brand" and "claims" features have the highest variable importance, as estimated through feature ablation. However, it must be pointed out that the process of feature ablation does not provide evaluations that are significantly distinguishing from the baseline metrics.

According to Section 4.1, the existing literature also argued that claims on meat substitute products retain the power to motivate buying intentions. Additionally, consumers nowadays care very much about the company or brand's reputation. The results that this study has produced are aligned with the findings of existing literature. Nonetheless, the product features that express less predicting power, such as the "format_t" variable, which contradicts the fact that was brought up in the literature. The same goes for product price - it shows mediocre predication ability in this study, but a key role in local consumer's decision-making process for plant-based products in the existing literature.

Limitations of the predicting importance of product features object in the method of feature ablation. It is mainly because there might be a chance that non-significant variations appear due to randomness and non-linearity. This scenario may happen because of the fact that some features interact with each other, and it will not be representative of feature independence.

7.3 *Recommendations for Future Work*

In future work, it is recommended that other categorical encoding methods like target encoder should be investigated further and see whether there is a growth in the metrics performance. Furthermore, other machine learning models can be explored when more data becomes available, especially of more balanced classes. Moreover, feature selection can be considered to reach more concrete results, which could also initiate a more practical way for businesses to fulfill later. Due to the limitation in terms of time and computation capacity, the grid search for hyper-parameter tuning involves two to four sets of different parameters on average. In future work, researchers can further develop more sets of parameters for the grid search process on each model.

8 CONCLUSION

The main research question for the present study is: *"What is the impact of product features on the geographical market decision of plant-based meat companies when entering a new market?"*. Two sub-questions are devised in order to navigate the present research. The first sub-question is: *"Which machine learning algorithm is a good fit for discovering the relationship between product features?"*. Five machine learning algorithms are resolved and modified by hyper-parameter tuning. The five models include NB, KNN, RFC, SVM and XGB. Overall, the performance of the NB algorithm exceeds other models that are generally considered more sophisticated with a balanced accuracy score of 0.48 and weighted F1 score of 0.61. Results from this study demonstrate that NB indeed is adequate in handling data that is high-dimensional and is presumed strong feature independence. The second sub-question is: *"Which product features have the most impact on the decision of entering a market?"*. Through the feature ablation technique, predictor features are ranked in a descending order attributed to each performing score. The product features of "company", "brand" and "claims" take over the top three features that hold the strongest perdition performance. Although some results are gained, these could also be elaborated by using feature selection method to achieve linearity and precision in the scientific point of view. When it comes to the societal relevance, meat substitute companies should pay more attention to the product features of company names, brands and claims indicated on packaging in order to succeed exporting to a new market.

Further research needs to investigate ways in terms of encoding techniques for individual product feature to further intensify classification predictions. Furthermore, more observations can be collected to reduce the class imbalance in the dataset. Due to the limitation of the feature ablation technique, feature selection, together with human intervention should be engaged. On the whole, the benefit for meat substitute companies can only reach the peak when the consumer's perception can be understood and taken action of, together with the advantage that machine learning models can bring.

REFERENCES

Alanís, P. J., de la Lama, G. C. M., Mariezcurrena-Berasain, M. A., Barbabosa-Pliego, A., Rayas-Amor, A. A., & Estévez-Moreno, L. X. (2022, 2). Sheep meat consumers in mexico: Understanding their perceptions, habits, preferences and market segments. *Meat Science*, *184*. doi: 10.1016/j.meatsci.2021.108705

Ampomah, E. K., Nyame, G., Qin, Z., Addo, P. C., Gyamfi, E. O., & Gyan, M. (2021, 6). Stock market prediction with gaussian naïve bayes machine learning algorithm. *Informatica (Slovenia)*, *45*, 243-256. doi: 10.31449/inf.v45i2.3407

Athey, S., Tibshirani, J., & Wager, S. (2019, 4). Generalized random forests. *Annals of Statistics*, *47*, 1179-1203. doi: 10.1214/18-AOS1709

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021, 3). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*, 1937-1967. doi: 10.1007/s10462-020-09896-5

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123-140. doi: 10.1007/bf00058655

Carlsson, F., Kataria, M., & Lampi, E. (2022, 3). How much does it take? willingness to switch to meat substitutes. *Ecological Economics*, *193*. doi: 10.1016/j.ecolecon.2021.107329

Carranza, C., Nolet, C., Pezij, M., & van der Ploeg, M. (2021, 2). Root zone soil moisture estimation with random forest. *Journal of Hydrology*, *593*. doi: 10.1016/j.jhydrol.2020.125840

Chen, J., Mai, W., Lian, X., Yang, M., Sun, Q., Gao, C., ... Chen, X. (2022). Ignoring encrypted protocols: Cross-layer prediction of video streaming qoe metrics. *Mobile Networks and Applications*. doi: 10.1007/s11036-021-01890-7

Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., ... Ren, H. (2020, 3). Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Research*, *171*. doi: 10.1016/j.watres.2019.115454

Chen, T., & Guestrin, C. (2016, 8). Xgboost: A scalable tree boosting system. In (Vol. 13-17-August-2016, p. 785-794). Association for Computing Machinery. doi: 10.1145/2939672.2939785

Chen, Y., & Zhang, Z. (2022, 6). Exploring public perceptions on alternative meat in china from social media data using transfer learning method. *Food Quality and Preference*, *98*. doi: 10.1016/j.foodqual.2022.104530

Chen, Z., & Fan, W. (2021, 8). A freeway travel time prediction method based on an xgboost model. *Sustainability (Switzerland)*, *13*. doi: 10.3390/su13158577

Chopra, A. B., & Dixit, V. S. (2021). Balanced accuracy of collaborative recommender system. In (Vol. 1270, p. 341-356). Springer Science and Business Media Deutschland GmbH. doi: 10.1007/978-981-15-8289-9_32

Collier, E. S., Oberrauter, L. M., Normann, A., Norman, C., Svensson, M., Niimi, J., & Bergman, P. (2021, 12). Identifying barriers to decreasing meat consumption and increasing acceptance of meat substitutes among swedish consumers. *Appetite*, *167*. doi: 10.1016/j.appet.2021.105643

Deng, Z., Zhu, X., Cheng, D., Zong, M., & Zhang, S. (2016, 6). Efficient knn classification algorithm for big data. *Neurocomputing*, *195*, 143-148. doi: 10.1016/j.neucom.2015.08.112

Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018, 4). *Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary* (Vol. 61). AI Access Foundation. doi: 10.1613/jair.1.11192

Gomes, H. M., Bifet, A., Read, J., Barddal, J. P., Enembreck, F., Pfharinger, B., ... Abdessalem, T. (2017, 10). Adaptive random forests for evolving data stream classification. *Machine Learning*, *106*, 1469-1495. doi: 10.1007/s10994-017-5642-8

He, H., Sun, M., Li, X., & Mensah, I. A. (2022, 4). A novel crude oil price trend prediction method: Machine learning classification algorithm based on multi-modal data features. *Energy*, *244*. doi: 10.1016/j.energy.2021.122706

He, S., Li, B., Peng, H., Xin, J., & Zhang, E. (2021). An effective cost-sensitive xgboost method for malicious urls detection in imbalanced dataset. *IEEE Access*, *9*, 93089-93096. doi: 10.1109/ACCESS.2021.3093094

Huang, S., Nianguang, C. A., Pacheco, P. P., Narandes, S., Wang, Y., & Wayne, X. U. (2018, 1). *Applications of support vector machine (svm) learning in cancer genomics* (Vol. 15). International Institute of Anticancer Research. doi: 10.21873/cgp.20063

Hussein, A. S., Li, T., Yohannese, C. W., & Bashir, K. (2019). A-smote: A new preprocessing approach for highly imbalanced datasets by improving smote. *International Journal of Computational Intelligence Systems*, *12*, 1412-1422. doi: 10.2991/ijcis.d.191114.002

Kantono, K., Hamid, N., Ma, Q., Chadha, D., & Oey, I. (2021, 9). Consumers' perception and purchase behaviour of meat in china. *Meat Science*, *179*. doi: 10.1016/j.meatsci.2021.108548

Kerslake, E., Kemper, J. A., & Conroy, D. (2022, 3). What's your beef with meat substitutes? exploring barriers and facilitators for meat substitutes in omnivores, vegetarians, and vegans. *Appetite*, *170*. doi: 10.1016/j.appet.2021.105864

Kim, M., Kang, Y., You, S. C., Park, H. D., Lee, S. S., Kim, T. H., ... Joung,

B. (2022, 12). Artificial intelligence predicts clinically relevant atrial high-rate episodes in patients with cardiac implantable electronic devices. *Scientific Reports*, *12*. doi: 10.1038/s41598-021-03914-4

Kim, S. M., & Park, M. J. (2020, 12). Evaluation of cross-national global market segmentation and strategy: The case of korean wave for asean countries. *Asia Pacific Management Review*, *25*, 207-215. doi: 10.1016/j.apmrv.2020.04.001

Lazzarini, G. A., Visschers, V. H., & Siegrist, M. (2017, 9). Our own country is best: Factors influencing consumers' sustainability perceptions of plant-based foods. *Food Quality and Preference*, *60*, 165-177. doi: 10.1016/j.foodqual.2017.04.008

Lemken, D., Spiller, A., & Schulze-Ehlers, B. (2019, 12). More room for legume – consumer acceptance of meat substitution with classic, processed and meat-resembling legume products. *Appetite*, *143*. doi: 10.1016/j.appet.2019.104412

Leong, W. C., Bahadori, A., Zhang, J., & Ahmad, Z. (2021). Prediction of water quality index (wqi) using support vector machine (svm) and least square-support vector machine (ls-svm). *International Journal of River Basin Management*, *19*, 149-156. doi: 10.1080/15715124.2019.1628030

Li, C., Zheng, X., Yang, Z., & Kuang, L. (2018). Predicting short-term electricity demand by combining the advantages of arma and xgboost in fog computing environment. *Wireless Communications and Mobile Computing*, *2018*. doi: 10.1155/2018/5018053

Li, H., Cao, Y., Li, S., Zhao, J., & Sun, Y. (2020, 5). Xgboost model and its application to personal credit evaluation. *IEEE Intelligent Systems*, *35*, 52-61. doi: 10.1109/MIS.2020.2972533

Meenakshi, M., & Geetika, G. (2014, 1). Survey on classification methods using weka. *International Journal of Computer Applications*, *86*, 16-19. doi: 10.5120/15085-3330

Mohammed, M., & Omar, N. (2018). Question classification based on bloom's taxonomy using enhanced tf-idf. *International Journal on Advanced Science, Engineering and Information Technology*, *8*, 1679-1685. doi: 10.18517/ijaseit.8.4-2.6835

Nalepa, J., & Kawulok, M. (2019, 8). *Selecting training sets for support vector machines: a review* (Vol. 52). Springer Netherlands. doi: 10.1007/s10462-017-9611-1

Pakravan, M., & Jahed, M. (2022, 3). Significant pathological voice discrimination by computing posterior distribution of balanced accuracy. *Biomedical Signal Processing and Control*, *73*. doi: 10.1016/j.bspc.2021.103410

Pisner, D. A., & Schnyer, D. M. (2019, 1). *Support vector machine.* Elsevier.

doi: 10.1016/B978-0-12-815739-8.00006-7

Sandøe, P., Hansen, H. O., Forkman, B., van Horne, P., Houe, H., de Jong, I. C., . . . Christensen, T. (2022, 2). Market driven initiatives can improve broiler welfare – a comparison across five european countries based on the benchmark method. *Poultry Science*, 101806. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0032579122001146 doi: 10.1016/j.psj.2022.101806

Sattari, M. T., Falsafian, K., Irvem, A., S, S., & Qasem, S. N. (2020, 1). Potential of kernel and tree-based machine-learning models for estimating missing data of rainfall. *Engineering Applications of Computational Fluid Mechanics*, *14*, 1078-1094. doi: 10.1080/19942060.2020.1803971

Schonlau, M., & Zou, R. Y. (2020, 3). The random forest algorithm for statistical learning. *Stata Journal*, *20*, 3-29. doi: 10.1177/1536867X20909688

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019, 11). *A comparison of random forest variable selection methods for classification prediction modeling* (Vol. 134). Elsevier Ltd. doi: 10.1016/j.eswa.2019.05.028

Sun, J., Du, W., & Shi, N. (2018, 5). A survey of knn algorithm. *Information Engineering and Applied Computing*, *1*. doi: 10.18063/ieac.v1i1.770

Syamsuddin, I., & Barukab, O. M. (2022, 3). Sukry: Suricata ids with enhanced knn algorithm on raspberry pi for classifying iot botnet attacks. *Electronics (Switzerland)*, *11*. doi: 10.3390/electronics11050737

Taşcı, E. (2019, 4). A meta-ensemble classifier approach: Random rotation forest. *Balkan Journal of Electrical and Computer Engineering*, 182-187. doi: 10.17694/bajece.502156

Trevethan, R. (2017, 11). Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. *Frontiers in Public Health*, *5*. doi: 10.3389/fpubh.2017.00307

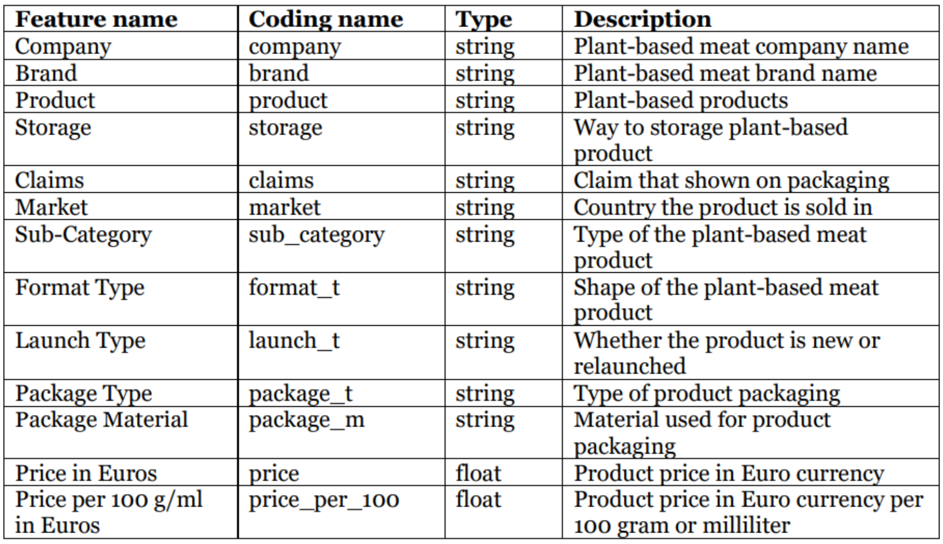
van den Bulk, L. M., Bouzembrak, Y., Gavai, A., Liu, N., van den Heuvel, L. J., & Marvin, H. J. (2022, 1). Automatic classification of literature in systematic reviews on food safety using machine learning. *Current Research in Food Science*, *5*, 84-95. doi: 10.1016/j.crfs.2021.12.010

Wang, R., Lu, S., & Li, Q. (2019, 8). Multi-criteria comprehensive study on predictive algorithm of hourly heating energy consumption for residential buildings. *Sustainable Cities and Society*, *49*. doi: 10.1016/j.scs.2019.101623

Xing, W., & Bei, Y. (2020). Medical health big data classification based on knn classification algorithm. *IEEE Access*, *8*, 28808-28819. doi: 10.1109/ACCESS.2019.2955754

Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017, 1). Learning k for knn classification. *ACM Transactions on Intelligent Systems and Technology*, *8*. doi: 10.1145/2990508

APPENDIX A



Feature name	**Coding name**	**Type**	**Description**
Company	company	string	Plant-based meat company name
Brand	brand	string	Plant-based meat brand name
Product	product	string	Plant-based products
Storage	storage	string	Way to storage plant-based product
Claims	claims	string	Claim that shown on packaging
Market	market	string	Country the product is sold in
Sub-Category	sub_category	string	Type of the plant-based meat product
Format Type	format_t	string	Shape of the plant-based meat product
Launch Type	launch_t	string	Whether the product is new or relaunched
Package Type	package_t	string	Type of product packaging
Package Material	package_m	string	Material used for product packaging
Price in Euros	price	float	Product price in Euro currency
Price per 100 g/ml in Euros	price_per_100	float	Product price in Euro currency per 100 gram or milliliter

Figure 7: Feature description of the plant-based meat product dataset

APPENDIX B

Table 8: Overall scoring performance on NB with feature ablation

	Feature Performance	
	Balanced accuracy	Weighted F_1 score
company	0.069	0.07
brand	0.083	0.068
claims	0.022	0.018
storage	0.009	0.002
sub_category	0.009	0.001
price	0	0.001
price_per_100	0	0
package_m	0.001	-0.002
package_t	0.008	-0.003
launch_t	-0.02	-0.003
format_t	-0.02	-0.007
product	-0.001	-0.009