Big Data Analytics on Food Products Around the World

Karthik Vegi
Indiana University Bloomington
College Mall Apartments
Bloomington, Indiana 47401
kvegi@iu.com

Nisha Chandwani
Indiana University Bloomington
Park Doral Apartments
Bloomington, Indiana 47408
nchandwa@iu.edu

ABSTRACT

Food is one of the basic necessities of human-being. It helps us gain energy to recharge our body to do the daily activities of moving, playing, and thinking. From being a cave man to producing a wide variety of foods, we have come a long way. The civilizations shaped the food habits of the world and there is a lot of variance in the food habits across countries. We analyze the *Open Food Facts* database that gathers information on food products from around the world to unearth some food habits of the world and we predict the food grade based on the nutrition facts of the food products.

KEYWORDS

i523, hid231, hid203, big data, food habits, food products, nutrition

1 INTRODUCTION

Open Food Facts is a non-profit initiative started by Stephane Gigandet and run by thousands of volunteers around the world. Any person around the world can contribute to the database by simply scanning a product using a mobile app which is made available to IOS and Android. This massive database of food products opens up a lot of opportunities to analyze the food products around the world and understand the food habits. We are particularly interested in the consumption of nutrients that come along with the food items across the world, the composition of different fat content, and the prediction of nutrition grade based on the nutrients.

2 FOOD ANALYSIS: IMPORTANCE AND RELATED WORK

In recent times, more and more companies try to market their food as low-fat or low-calories in order to fool consumers into buying their products. The increasing concern of public health has led to a significant interest in detecting the health-related properties of food products [2]. Thus, there is no question about the importance of analysis of the nutrition grade and food safety in today's world. The analysis of food requires more robust and efficient methodologies in order to ensure the quality and safety of the food products [2]. Previous methods based on the so-called wet-chemistry have now evolved into more powerful techniques which are used in the food laboratories. These methods provide a massive improvement in analytical accuracy thus expanding the limits of food applications [2]. The traditional methods of food analysis can be classified based on the underlying principle. Some of these categories are spectroscopic, biological, electrochemical, supercritical fluid chromatography [2]. All these techniques provide information about the sample under study and this information is derived from a specific physical-chemical interaction [2]. A different approach to analyzing and detecting the food quality is by using machine learning techniques. We will discuss one of these modern methods of food analysis which can be widely used across countries.

3 ANALYSIS OF NUTRIENTS IN FOOD

Fat is definitely a nutrient that the body needs and is an essential nutrient that aids in cell growth, helps with energy generation, maintaining body temperature, protect organs, help absorb other essential nutrients that aid in producing energy, improve blood cholesterol level, help reduce inflammation in case of injury, and help in storing energy that can be used for survival when you go without food for few days [1]. But we do need to keep a track of the consumption because anything that is remotely excess leads to a variety of serious health issues [1].

3.1 Dietary Fats

There are different types of fat fi?? some are good and some are bad and some needs to be taken within a certain limit [1].

3.1.1 *Saturated Fat.* More intake of saturated fats results in the cholesterol levels in the blood which increases the risk of heart-related diseases [1]. The American Heart Association suggests around 5 percent of daily calories from foods containing saturated fat [1]. Meat, cheese, and milk are some of the sources of saturated fat [1].

3.1.2 *Trans Fat.* Any type of trans fat whether it is natural or artificial is not good [1]. The reason why food manufacturers use trans-fat is that they are less expensive, can be produced artificially, easy to use with other ingredients, last for a long time and also aid in improving the taste of the food [1]. Trans fats raise the bad fat levels and decrease the good fat levels [1]. The American Heart suggests to completely cut off trans-fat from the diet [1].

3.1.3 *Monounsaturated Fat.* Monounsaturated fats have a good effect on the body when taken within limit [1]. They help reduce the bad cholesterol levels in the blood and thereby decrease the risk of heart diseases [1]. They also help in gaining vitamin E which is a good nutrient that acts as antioxidant [1]. Olive oil, avocados, and sesame oil are some of the sources of monounsaturated fats [1].

3.1.4 *Polyunsaturated Fat.* Polyunsaturated fats have a good effect on the body when taken within limit [1]. They help reduce the bad cholesterol levels in the blood and thereby decrease the risk of heart diseases [1]. They also provide some nutrients that are essential for the body [1]. Soybean oil and sunflower oil are some of the sources of polyunsaturated fats [1].

3.2 Data Cleaning and Transformation

To make the analysis more interesting, the top 20 countries with most value counts for the attributes have been considered. The countries with names combined with other countries were also cleaned in the process. The data was analyzed for missing values and the attributes with more than 60 percent missing values were removed from the analysis to add consistency. Only the columns that are meaningful in the analysis were retained and the rest were removed from further analysis.

We then display the top 5 countries as a pie-chart and the 5 countries are namely United States, France, Switzerland, Germany, and Spain as shown in Figure 1.

[Figure 1 about here.]

We then impute all the null values with zeroes and we then check the dietary fat content in the foods and check the top countries with fat content using a histogram. The analysis with respect to the fat countries is as follows

3.3 Fat Content

The top 5 countries with most fat content in the food items are Serbia, United States, Switzerland, Germany, and Sweden as shown in Figure 2.

[Figure 2 about here.]

The top 5 countries with most saturated fat content in the food items are Serbia, United States, Germany, France and Switzerland as shown in Figure 3

[Figure 3 about here.]

The top 5 countries with most trans-fat content in the food items are United States, Brazil, Canada, Australia, Russia, and Serbia as shown in Figure 4.

[Figure 4 about here.]

The top 5 countries with most cholesterol content in the food items are United States, Canada, Portugal, Brazil, France, and Italy as shown in Figure 5.

[Figure 5 about here.]

3.4 Sugar and Salt Content

Although the body needs sugar, high intake of artificial and processed sugar is bad for health as it does not add any nutrients but only adds calories [5]. It is always better to rely on the natural sugar that comes with fruits and milk [5]. Artificial sugars tooth decay and diabetes [5]. Just as fat, sodium which is the main source of iodine is essential for health although its intake should be within limit [5]. Increase in intake of salt leads to blood pressure and has an effect on the heart [5].

The top 5 countries with most sugar content in the food items are United States, Serbia, Switzerland, France, and Sweden as shown in Figure 6.

[Figure 6 about here.]

The top 5 countries with most sodium content in the food items are United States, Hungary, Serbia, Sweden, and France as shown in Figure 7.

[Figure 7 about here.]

4 NUTRITION GRADE LABELLING SYSTEM

France recently took a decision to implement a nutri-score system which will use a color coding mechanism to label the food products that will help consumers know the nutrition grade of the product [3]. The World Health Organization regional office for Europe as a part of its 5-year action plan from 2015-2020 recommends a labeling mechanism for the consumers to know about the quality of the food products at a first glance [3]. This will not only make it easier for the consumers to pick healthier options but it will also regulate food manufacturers to resort to healthier ingredients instead of going for low cost artificial or less healthy ingredients [3].

France after the United Kingdom became the second country to implement this system to indicate the main ingredients like fat, salt and sugar content in the food items [3]. France made use of an evidence-based system to study different labeling systems to arrive at the best one [3]. By implementing this system, the World Health Organization will keep a check on the growing number of diet-related diseases in the Europe region [3]. Europe being the largest consumer of cheese wants to regulate the ingredients that go into the manufacturing process so that people are well informed about their food choices [3].

4.1 Nutrition Grade Prediction as a Big Data Problem

We build a predictive classification model to predict the food nutrition grade based on the ingredients of the food. The goal is to apply various machine learning algorithms to the problem at hand, measure the prediction accuracy to compare and contrast the different algorithms and arrive at the best algorithm that suits the given data and the problem. This problem can be solved using Big Data and Machine Learning techniques given the size and the complexity of the data.

5 MACHINE LEARNING

Machine Learning is a field in which we train computers in a way that they can learn from the input data [6]. The ideology is that computers use the training data that is made available to them, learn from it, build a model and use this experience to build knowledge that can be applied on new unseen data [6]. A wonderful example to demonstrate machine learning is the application to detect spam emails where the machine builds knowledge from previously seen emails which are marked as spam, checks new emails to see if they match the historic spam emails and label them as spam or non-spam [6].

5.1 Types of Machine Learning Algorithms

There are primarily two types of machine learning algorithms, descriptive models and predictive models [6]. A *Descriptive Model* is described as the analysis done and insights gained from slicing and dicing the data in new and interesting ways [6]. One example of a descriptive model is pattern discovery that is often used in market basket analysis where transnational purchase details are analyzed [6]. A *Predictive Model* on the other hand involves predicting one value using one or more variables [6]. The learning algorithms tried to build a model that captures the relationship between a response variable and the independent variables [8].

5.2 Types of Learning

Unsupervised Learning is the process where there is no explicit training data to learn from, so there is simply no mechanism where the machine can learn from previously available data [6]. The same email example can be looked at in a different way where we now want to do anomaly detection in emails [6]. Here the main goal is to detect unusual messages from the bunch of messages and we do not have experience of previous data [6].

Supervised Learning in contrast is the process of gaining knowledge or expertise from the training data which can be applied to future unseen data [6]. Here the model is first trained by using a bulk of training examples and this model is applied to testing data to measure the accuracy [6]. The variable that we need to predict is identified which is called the response variable and the variables that are used to predict the response variables, called the predictor variables are identified [6]. If the existing variables are not sufficiently giving the accuracy that is expected, a method called feature engineering is done where new variables are derived by combining existing variables [6].

6 PREDICTION ANALYSIS

Prediction analysis is the process of working on a large dataset using a combination of statistical, data mining and machine learning algorithms to predict the outcome based on past data [6]. There are primarily two types of prediction analysis in machine learning, namely regression and classification [8]. In regression, we try to predict a continuous variable from the predictor variables [8]. A good example of regression is to predict the housing prices from different parameters like the year of construction, location, amenities, number of bedrooms etc [8]. Here the response variable is continuous and it is not predefined [8]. Classification, on the other hand, tries to predict a categorical variable in which we assign each record with a predefined label or a class [8].

Classification is the task of assigning each data record to a predefined class [8]. In machine learning, classification is categorized as a supervised learning technique [8]. This problem has applications in various fields like spam detection, medical applications, astronomy, and banking to identify fraudulent transactions from genuine transactions [8]. It is the task of coming up with a model which is essentially a function that maps every data record to a class label [8].

The task at hand is a classification problem since we are trying to predict the food nutrition grade of the products based on the ingredients that go into the product. For this problem, we are considering only the data for the country France, since the nutrition grade is available for most food products from the country. Another reason is that France is the first country in the region to come up with the idea of adding a color-coded label to the food products mentioning the nutrition grade. In the subsequent sections, we discuss the machine learning techniques used to solve this problem.

6.1 K Nearest Neighbors

6.1.1 Overview. Some of the classification algorithms in machine learning work on the principle of eager learning that involves a two-step process where first a model is built from the training data and the model is applied on testing data [8]. In contrast, K nearest neighbors is a lazy learning algorithm where the process of modeling the training data is not done until the test examples are classified [8]. *Rote Classifier* is a good example of lazy learning algorithm which memorizes the entire training data to perform classification but has the drawback of not being able to map every test example against the training example [8]. K nearest neighbors algorithm overcomes this drawback by finding all the records that are closest or nearest to the training records [8].

The nearest neighbor puts each attribute list as a data point in the n-dimensional space, given n the number of attributes [8]. Once we have the training examples, we take each test example and compute its distance to the training example classes and assign a class label [8]. Any of the popular distance measures among Euclidean distance, Manhattan distance, Minkowski distance and Mahalanobis distance can be used [8]. The k denotes the k closest points to the test example [8]. Figure 8 shows the algorithm [8].

[Figure 8 about here.]

6.1.2 Support in Python. KNeighborsClassifier is available in the scikit learn python library.

6.2 Logistic Regression

Logistic regression or logit regression is a special type of regression analysis where the response variable that we need to predict is a categorical variable [8]. Typically, logistic regression models the response variable to take two values, 1 or 0, pass or fail, win or lose [8]. Logistic regression that takes more than two values for the response variable is called multinomial logistic regression [8]. Here the probability of the response variable to take a categorical value is modeled as a function of the predictor variables [8].

Like a lot of machine learning algorithms, logistic regression works by making a lot of assumptions which should be taken care as a part of the data cleaning and transformation process [6]. It does not assume a linear relationship between the response variables and the predictor variables [6]. Since it applies a log transformation on the predicted probabilities, it can handle a variety of relationship between the predictor variables [6]. If the predictor variables are multivariate normal, the algorithm achieves the best result although it works even if they are not [6]. The stepwise method must be used in the logistic regression to ensure that we are neither overfitting nor underfitting the data [6]. A very important assumption to be noted in logistic regression is that each attribute list must be independent, in the sense, the data records must not be derived from a before-after setup experiment [6]. It also requires a decently large sample size to work on [6].

6.2.1 Support for Python. LogisticRegression is available in the scikit learn python library.

6.3 Random Forest Classifier

Random forest is an ensemble classification algorithm which is very powerful [8]. Ensemble method is a special process to improve the accuracy of the prediction [8]. The classification algorithms we have seen so far predict the response variable using a single classifier on the test data but ensemble methods use multiple classifiers in tandem and aggregate the predictions to boost the accuracy by a huge margin [8]. Using a combination method, the ensemble method derives a set of base classifiers from the training data and on each iteration takes a vote of all the base classifiers to arrive at a result [8].

Random forest is an ensemble method which works very well for classification problems [8]. It combines the predictions made by multiple classifiers where each classifier independently works on the training data and casts its vote [8]. Unlike methods like AdaBoost which generates values based on independent random vectors using a varied probability distribution, random forest generates values based on fixed probability distribution [8].

6.3.1 Rationale for Random Forest. Consider an example, where we have 25 base classifiers and each base classifier has an error rate of 0.35 [8]. As discussed, the random forest takes the majority vote given by the base classifiers [8]. The model makes a wrong prediction if half or more base classifiers predict inaccurately. The accuracy is improved with an error rate of 0.06 which is far better than using just a single classifier [8].

6.3.2 Support for Python. RandomForestClassifier is available in the scikit learn python library.

7 EXPERIMENTS AND RESULTS

In this section, we will introduce the algorithm along with the details of experiments and methodology for predicting the nutrition grade of food products in France.

7.1 Algorithm

The problem at hand is to correctly identify the nutrition grade of the food item. The possible labels are, *a* to *e*, with *a* being the best and *e* being the worst grade for a food item. For this task, we have used machine learning techniques that help in predicting the label of each food item. Before getting into the details of each step of the method, we first present a concise version of the algorithm used for this task:

(1) Select all the records for the country, France. Drop records where nutrition grade is not populated.
(2) Separate the predictors from the response variable in order to perform data cleaning and data transformation steps.
(3) Check for missing values in the predictors obtained in the step above. Drop columns with more than 60% missing values.
(4) Impute the missing values with 0 for remaining columns.
(5) After imputing the missing values, standardize all the numerical predictors using the standard scaler.
(6) Check for the correlation between different numerical predictors. Drop one predictor from each pair of predictors that show high correlation.
(7) Combine the pre-processed predictors and the response variable in a single data frame.
(8) Divide the data obtained in step above into training and test data using stratified sampling.
(9) Train different classifiers on the training data and check the performance of each classifier on the test data.

7.2 Data set

For the classification problem, we selected the records for country France.

Number of examples: 123,961
Number of variables: 12
Response variables: *Nutrition Grade*
Predictor variables: *Energy per 100g, Fat per 100g, Saturated Fat per 100g, Carbohydrates per 100g, Sugars per 100g, Fiber per 100g, Proteins per 100g, Salt per 100g, Trans-fat per 100g, Sodium per 100g*

7.3 Python Packages Used

The following Python packages were used to solve the classification problem:

- Pandas: Provides high-performance data structures for data analysis and data munging
- Matplotlib: Plotting library that helps to embed plots into applications using GUI
- Seaborn: Visualization package based on matplotlib used for drawing high-level statistical graphics
- Scikit-learn: Toolbox with solid implementation of machine learning and other algorithms
- Scipy: Package that supports scientific computing with modules for linear algebra and integration

7.4 Data Cleaning

7.4.1 Step 1: Data Sparsity. Data sparsity refers to the situation where a lot of attributes have missing values which is an advantage in some cases because you only need to store and analyze the data that is available to you and save on computation time and storage [8]. We first check the data value counts for each country. United States, France, Switzerland, Germany, and Spain come as the top 5 countries with most data. Since the food nutrition grade was implemented in France, it has most products for which nutrition grade is labeled. So for this classification problem, we use the food data from France for analysis.

7.4.2 Step 2: Handling Missing Values. Missing values is a common scenario and they can be handled in different ways. You could choose to eliminate the data objects with missing values but at the expense of missing some critical analysis [8]. Estimating the missing values is also a good way to handle them, especially when the data comes from time series etc, where you could possibly interpolate the missing values from the ones that are closer to it [8]. Ignoring the missing values is another technique which can be applied to tasks like clustering where the similarity can be calculated using the attributes other than the missing ones [8].

The data set was first analyzed to check the missing values in all the columns. The threshold limit has been set at 60 percent. All the columns with missing values more than 60 percent were removed from the analysis to make the result more consistent. Once the columns were removed, the data set has to be re-indexed to maintain the order. Only the columns that are important for the prediction task have been retained from the original dataset. In this case, all the ingredients which are primarily the predictor variables were included. The missing values in the response variable also need to be taken care of. Removing the records with missing values for the response variable proved to be the best option for trying out various things.

Imputation was used to handle the null values in the predictor variables. Imputation can be done in a variety of ways, for example, replacing the missing values with zero or imputing the missing values for numerical columns with the mean and the categorical columns with the mode. Since all the predictor variables have numeric values, all the null values have been replaced with zero. To ensure that the imputation process has been done correctly, the sum of missing values is calculated since post-imputation, this sum should be zero.

7.4.3 Step 3: Outlier Treatment. Outliers are data objects with quite distinct characteristics from the other data records [8]. There is a considerable difference between anomalies and outliers, where anomalies refer to data records that have bad data, which is noise and need to be ignored, anomalies often contain interesting aspects and can lead to some good analysis [8]. In applications like *Fraud Detection*, anomalies could be of utmost importance [8]. The outliers in the data have been looked at by using box plots and have been handled as a part of the data cleaning process.

7.5 Exploratory Data Analysis

For exploratory data analysis, we used the Seaborn package along with Matplotlib for visualizations. The measure of spread, that is the range and variance of the values, is a good way to understand the different aspects of the predictor variables. Box-plots are a method of visualization to look at the distribution of values for a numerical attribute [8]. The box plots show the percentiles where the lower and upper ends of the box indicate 25^{th} and 75^{th} percentile, the line inside the box indicates the 50^{th} percentile, the tails indicate the 10^{th} and 90^{th} percentile respectively [8].

7.5.1 Bi-variate box-plots. Bi-variate box-plots go beyond univariate box plots by showing the relationship between the predictor variable and the response variable [8]. We look at the bi-variate box-plots for each of the important predictor variables namely, saturated fat, polyunsaturated fat, sugars and salt and the response variable, nutrition grade. Figure 9 shows the bi-variate box plots.

[Figure 9 about here.]

By looking at the box plots, we can understand some important aspects of how the response variable is related to the predictor variables. We see that as the average saturated fat content increases, the food grade decreases and as the average polyunsaturated fat content increases the nutrition grade is better. When the sugar levels increase, the health quotient of the food comes down. The energy levels behave in an interesting manner where the energy for the nutrition grade A is higher whereas in general, the average energy level slightly increases with the decreasing nutrition grade. While increase in energy does not necessarily imply that the nutrition quality is high, as there are a lot of instant energy foods that have a lot of additives, but they are often rated low when it comes to health.

7.5.2 Correlation. Correlation between data objects is the measure of the linear relationship between the attributes of the object that are continuous variables [8]. Correlation analysis is the process of finding of the correlations between the different predictor variables and identify high collinearity problem [6]. The relationship could be either linear or non-linear based on the given data [8]. The correlation coefficient can range anywhere between -1 and 1, where 1 indicates a very high positive correlation and -1 indicates a very high negative correlation [6]. Correlation plot visually shows the correlation coefficient between the variables in a nicely laid out plot. Figure 10 shows the correlation plot.

[Figure 10 about here.]

By looking at the correlation plot, we can see that sugars, fat, energy are positively correlated with the nutrition grade. This indicates that these variables will play an important role in the prediction algorithm. However, sodium and salt are highly correlated with each other and this may lead to collinearity problem if not handled. Collinearity is the state where the independent variables are highly correlated with each other which can add a lot of noise to the data [7]. Some of the problems because of collinearity are that the regression coefficients may not be estimated correctly. Also, collinearity makes it very difficult to explain the response variables using the predictor variables [7]. So we remove sodium from the predictor variables and proceed to the next step.

7.5.3 Data Transformation. Data transformation refers to the transformation that is applied to the variables [8]. For each data object, we apply a transformation function to all the attributes of the object to ensure that the attributes do not have a lot of variance in the data [8]. This process is also called standardization since we are applying a standard function to make sure all the attributes fall within a given range [8]. There are different methods that can be applied to achieve scaling namely log transformation, absolute value, square root transformation [8].

We use the method called normalization where all the values fall in the range, 0 to 1. To achieve this, we use the prepossessing package from sklearn which provides utility functions and transformer classes to change raw data into a standard representation. A lot of machine learning algorithms work well on standardized data. If some of the variables have extreme values, they might dominate the model function and might disturb the estimation parameter. Thus, for such extreme values, standardization helps achieve better results.

On scaling the data, there was a massive improvement in the prediction accuracy of the algorithms, implemented for this task. Thus, this proves the importance of data standardization with respect to machine learning algorithms.

7.6 Data Sampling

In a supervised machine learning approach, the model is trained on one sample of the data and later tested on a different sample of the data. Thus, in order to test the performance of the nutrition grade classifier, the data for the country France was divided into two samples, training and testing. There are various ways to achieve this split or sampling of the data. Some of these sampling methods are:

- Simple Random Sampling: This is one of the simplest sampling techniques. In this technique, every data point has an equal chance of being selected. In other words, it works similar to a lottery system where every outcome has an equal probability. The biggest advantage of this technique is the ease of implementation and its unbiased nature while generating the sample. However, random sampling might not always result in a sample that can represent the true population. It generally works well when we have huge data to sample from.
- Stratified Sampling: This technique is a more sophisticated method of sampling data. Stratified sampling generates a sample such that the proportion of each class in the sample is same as that in the true population. In this technique, the entire population is divided into groups or strata. The next step is to randomly select data points from each stratum such that the final sample has the same proportion for each stratum as that present in the true population. Thus, the sample generated by this technique is a good representative of the true population. Stratified sampling is a very useful technique when the classes in the data are highly imbalanced.

For our classifier, we chose to divide the data for France into training and test samples using stratified sampling technique. The strata or groups were created based on the response variable, i.e., food grade. This ensured that the training and test data had the same proportion of each food grade.

7.7 Data Modeling

Once the data was divided into training and test data, the next step was to train different classifiers and tune their respective parameters for better accuracy. We implemented three different models for classifying the food grade. Each of these models along with their parameters is:

- K Nearest Neighbors (kNN): For kNN, the grade of a food item in test data is classified by first finding the k most similar food items in the training data. It then takes the vote (food grade label) from each of these neighbors and based on the majority vote, the food item from the test data is assigned a food grade. Thus, one of the most important parameter for kNN is k, i.e., the number of neighbors to consider from the training data. We tried different k values and found that $k = 3$ gives the best accuracy.
- Logistic Regression: For logistic regression, one of the important parameters is the penalty. This parameter specifies the kind of regularization to be applied. This parameter can take two possible values, l_1 regularization and l_2 regularization. Both these values penalize high magnitude of the coefficients of the predictors in order to prevent the model from over-fitting. For our model, we have used l_2 regularization as it works well even in the presence of highly correlated features.
- Random Forest: For the random forest, there are many parameters, such as the number of trees in the forest, the maximum depth of the trees, maximum number of features to consider at each split, the minimum number of samples required in a sub-tree to qualify for a further split, the minimum number of samples required to qualify as a leaf node, etc. For our data, we have kept most of the parameters at their default values, except for, the number of estimators or trees in the forest. We have set this value to 100, as the classifier resulted in very high accuracy with 100 trees in the forest.

7.8 Evaluation Metrics and Results

There are various evaluation metrics for assessing the performance of classifiers. Some of these evaluation metrics are [4]:

- Accuracy: This metric gives the proportion of the total number of correctly classified instances
- Precision: This gives the proportion of the true positive instances from the total instances classified as positive
- Recall: This gives the proportion of the positive instances that are correctly classified
- F-Measure: This gives the harmonic mean between precision and the recall values
- Confusion Matrix: This is a useful way of checking the accuracy of the classifier. It clearly shows the number of instances correctly classified for each label. Thus, if we know that the classes in the data are not well-balanced, it's always a good idea to check the confusion matrix along with accuracy. Consider a case where 95% of the instances belong to class A and only 5% of the instances belong to class B. If a classifier is trained on a dataset with such imbalance, there is a high chance that the classifier would return label A for each test instance. The classifier would still be able to correctly classify 95% of the test instances resulting in 95% accuracy. This is a case where accuracy can be misleading and thus a quick look at the confusion

matrix can help understand the problem with the classifier. For such a case, the confusion matrix will clearly show that all the instances of the minority class, B, have been misclassified.

For our model, we used accuracy as well as confusion matrix for evaluating the results. The confusion matrix did not show any serious issues for any of the classifiers. The accuracy for each of the three classifiers was:

(1) Logistic Regression: With l_2 penalty, the accuracy of logistic regression was 78.9%. Figure 11 shows the confusion matrix.

[Figure 11 about here.]

(2) K Nearest Neighbors: With k as 3, the accuracy of kNN was 95.74%. Figure 12 shows the confusion matrix.

[Figure 12 about here.]

(3) Random Forest: With a number of trees as 100, the accuracy of random forest classifier was 99.68%. Figure 13 shows the confusion matrix.

[Figure 13 about here.]

Thus, we obtained the best results with Random Forest classifier.

8 CONCLUSION

Analysis of food content is very important in today's world as most of the companies try to fool consumers by labeling their product as low-fat. It's important for the consumers to know the true nutrition grade while purchasing any food item. Thus, we analyzed the nutrition grade based on the composition of various components of the food items. We developed a model that labels a food item purely on the basis of its nutrients, thus eliminating any bias, such as, the production company or the brand name. For accurate labeling, we applied different data cleaning and data transformation techniques. With this transformed data, we tried various machine learning models. We got the best results using random forest classifier which was able to accurately label 99% of the food products. Since the model is trained only for France, as part of future work, we can try and scale our model for different countries. However, to achieve similar results for other countries, we need to collect more data. The current data has many missing values for countries other than France. Once we collect enough data for these countries, we can also try and implement more sophisticated models like neural networks in future.

ACKNOWLEDGMENTS

This project was undertaken as a part of the course objective for I523: Big Data Applications and Analytics at Indiana University, Bloomington. We would like to thank Dr. Gregor von Laszewski and all the TAs for their help, support, and suggestions.

A WORK BREAKDOWN

Dataset identification: Karthik Vegi, Nisha Chandwani: work equally split between.

Requirement Gathering: Karthik Vegi, Nisha Chandwani: work equally split between.

Learning Machine Learning Concepts: Karthik Vegi, Nisha Chandwani: work equally split between.

Data analysis and implementation of the Logistic Regression: Karthik Vegi.

K nearest neighbors and Random Forest algorithms: Nisha Chandwani

Writing the project report: Karthik Vegi, Nisha Chandwani: work equally split between.

REFERENCES

[1] American Heart Association. 2017. Dietary Fats. Webpage. (March 2017). https://healthyforgood.heart.org/eat-smart/articles/dietary-fats

[2] Alejandro Cifuentes. 2012. Food analysis: present, future, and foodomics. *ISRN Analytical Chemistry* 2012 (2012), 16.

[3] World Health Organization Europe. 2017. Labelling systems to guide consumers to healthier options. Webpage. (March 2017). http://www.euro.who.int/en/countries/france/news/news/2017/03/france-becomes-one-of-the-first-countries-in-region-to-recommend-colour-coded-front-of-pack-

[4] M Hossin and MN Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process* 5, 2 (2015), 1.

[5] Healthy Eating SFGate. 2017. Recommended Daily Allowances of Fats, Sugars, Sodium for Adults. Webpage. (2017). http://healthyeating.sfgate.com/recommended-daily-allowances-fats-sugars-sodium-adults-2976.html

[6] Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, New York, USA.

[7] Statistics Solutions. 2017. Multicollinearity. Webpage. (March 2017). http://www.statisticssolutions.com/multicollinearity/

[8] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2006. *Introduction to Data Mining.* Pearson, Boston, USA.

[9] Karthik Vegi and Nisha Chandwani. 2017. Code base - Analysis on food products around the world. github. (Dec. 2017). https://github.com/bigdata-i523/hid231/tree/master/project/code

List of Figures

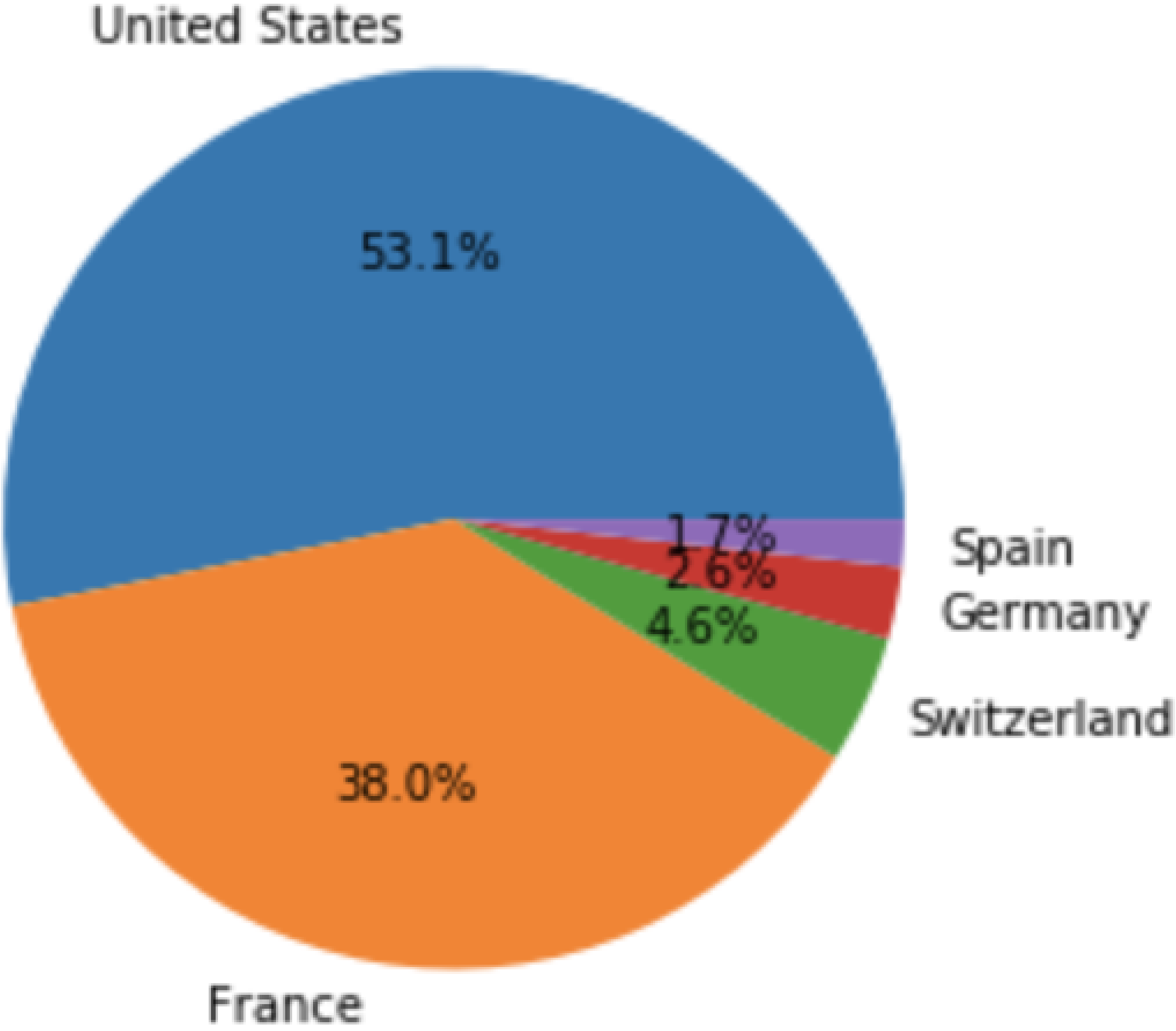


Figure 1: Top 5 countries [9]

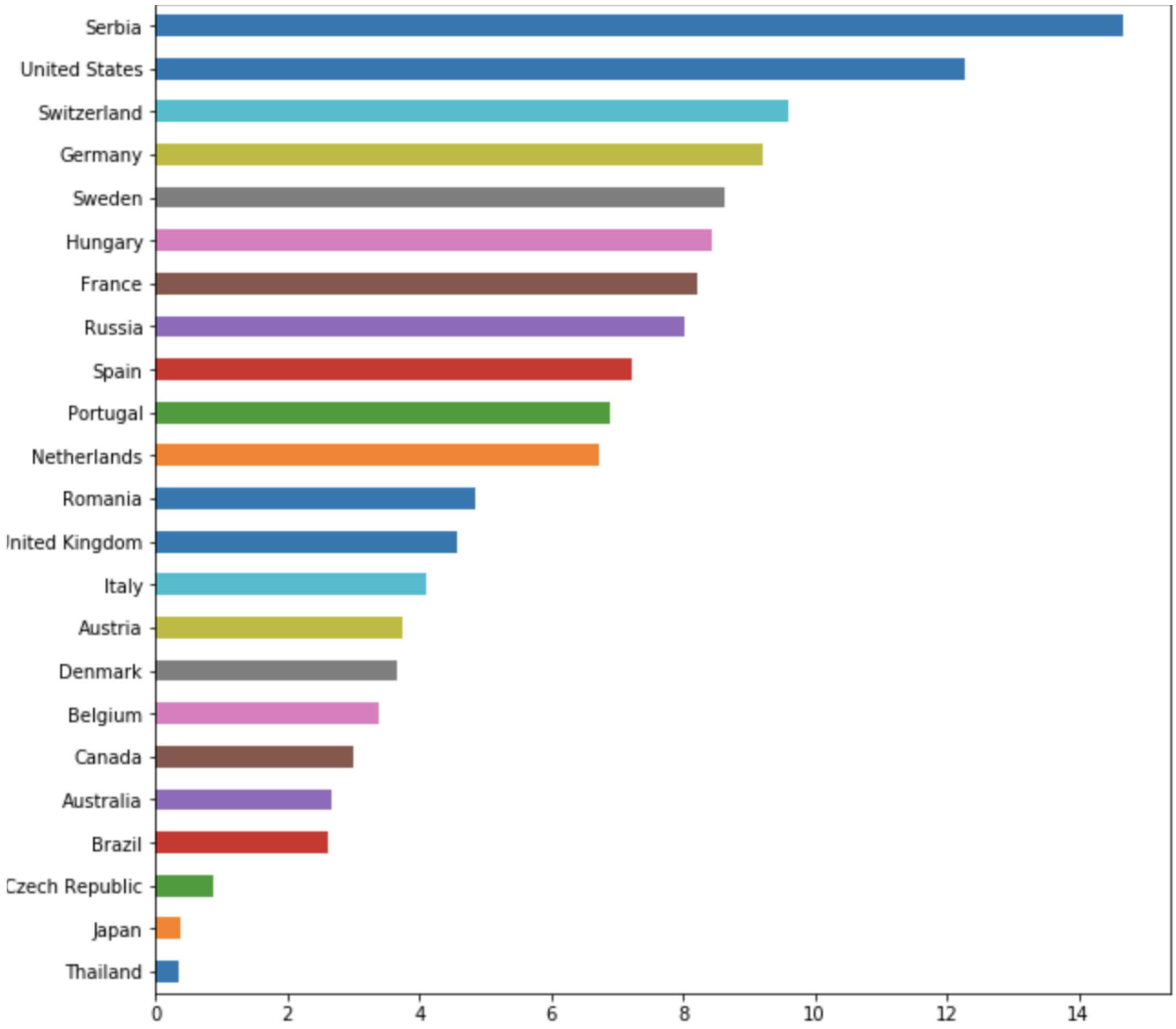


Figure 2: Top 5 countries with most fat content [9]

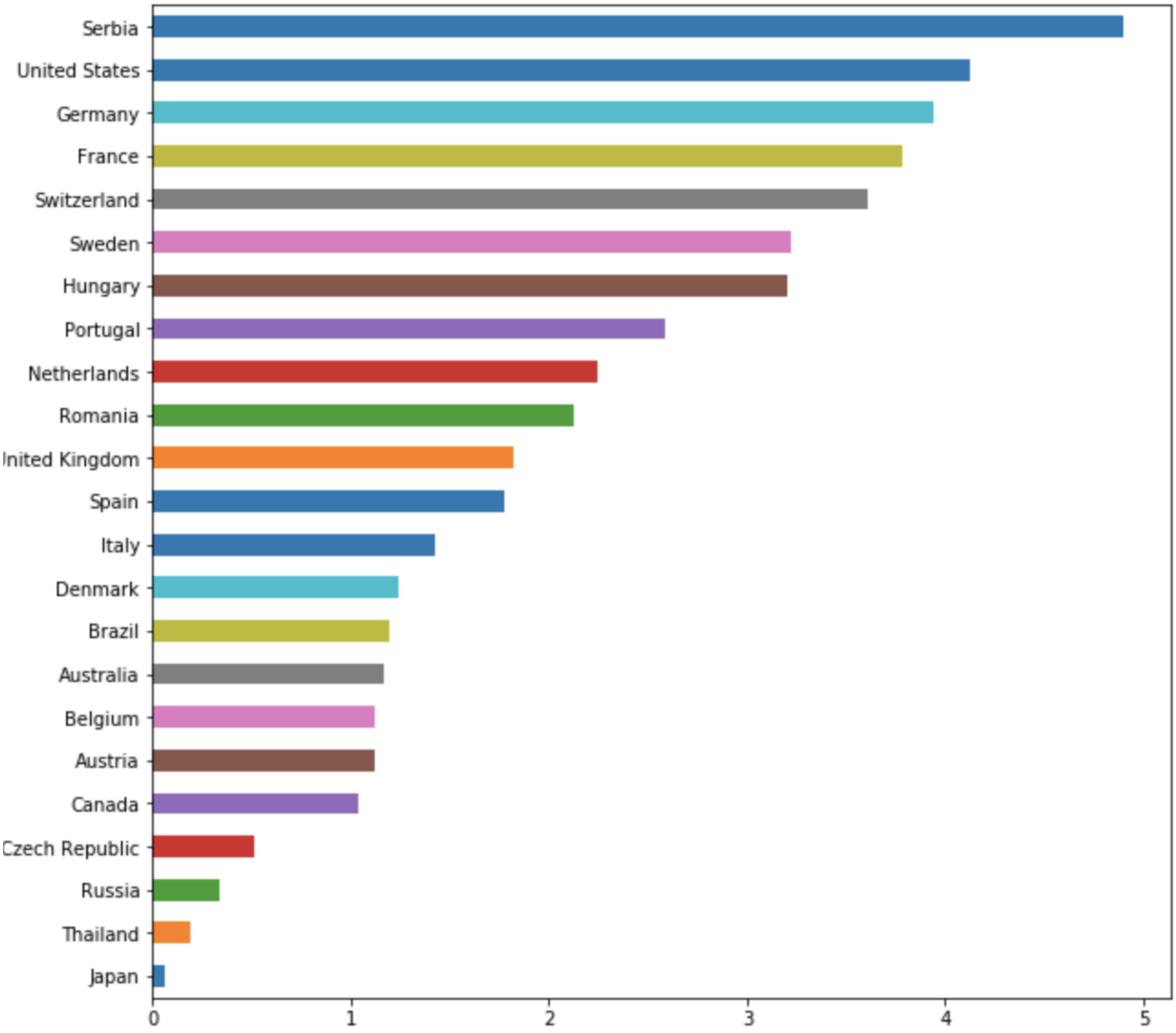


Figure 3: Top 5 countries with most saturated fat content [9]

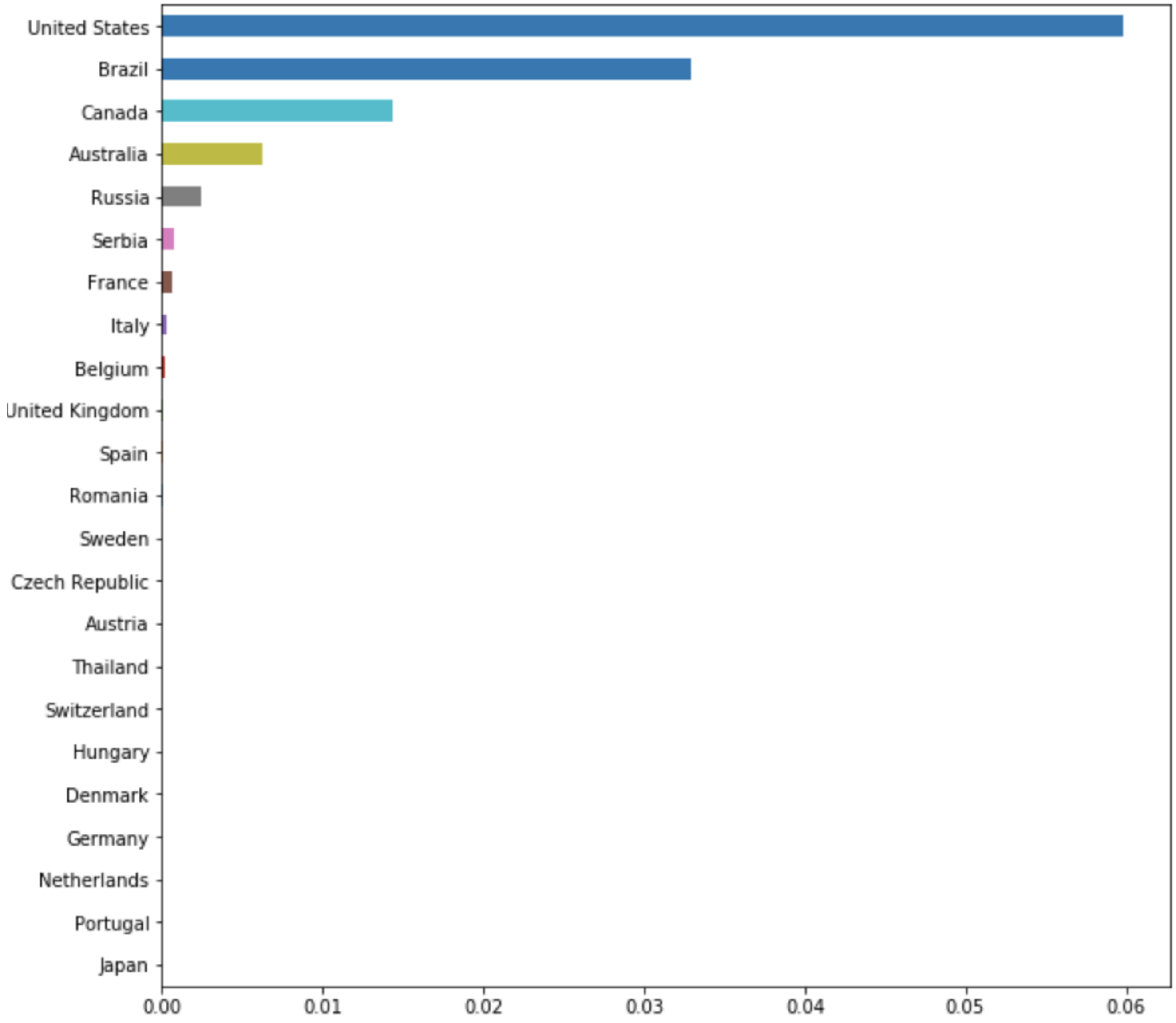


Figure 4: Top 5 countries with most trans-fat content [9]

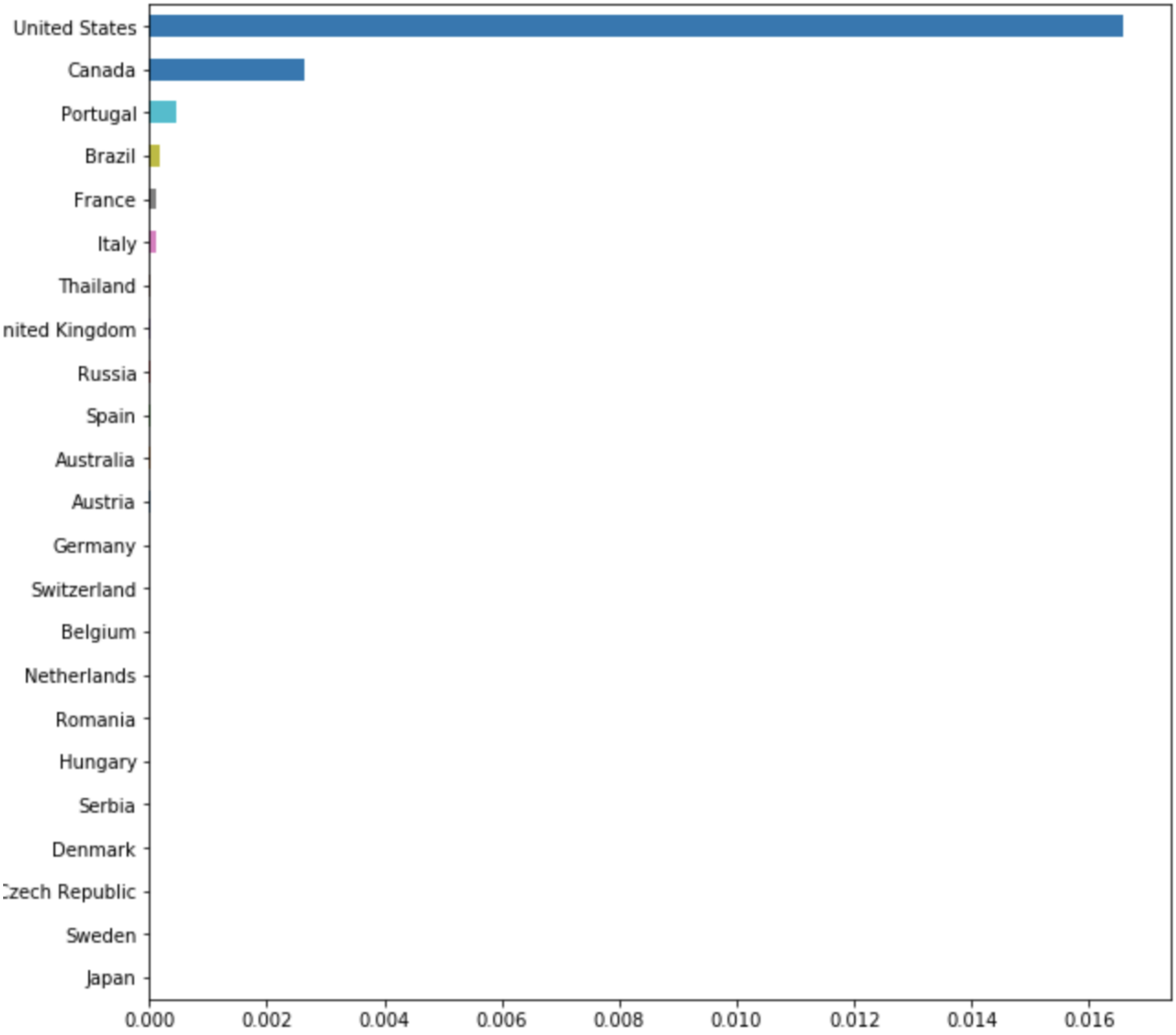


Figure 5: Top 5 countries with most cholesterol content [9]

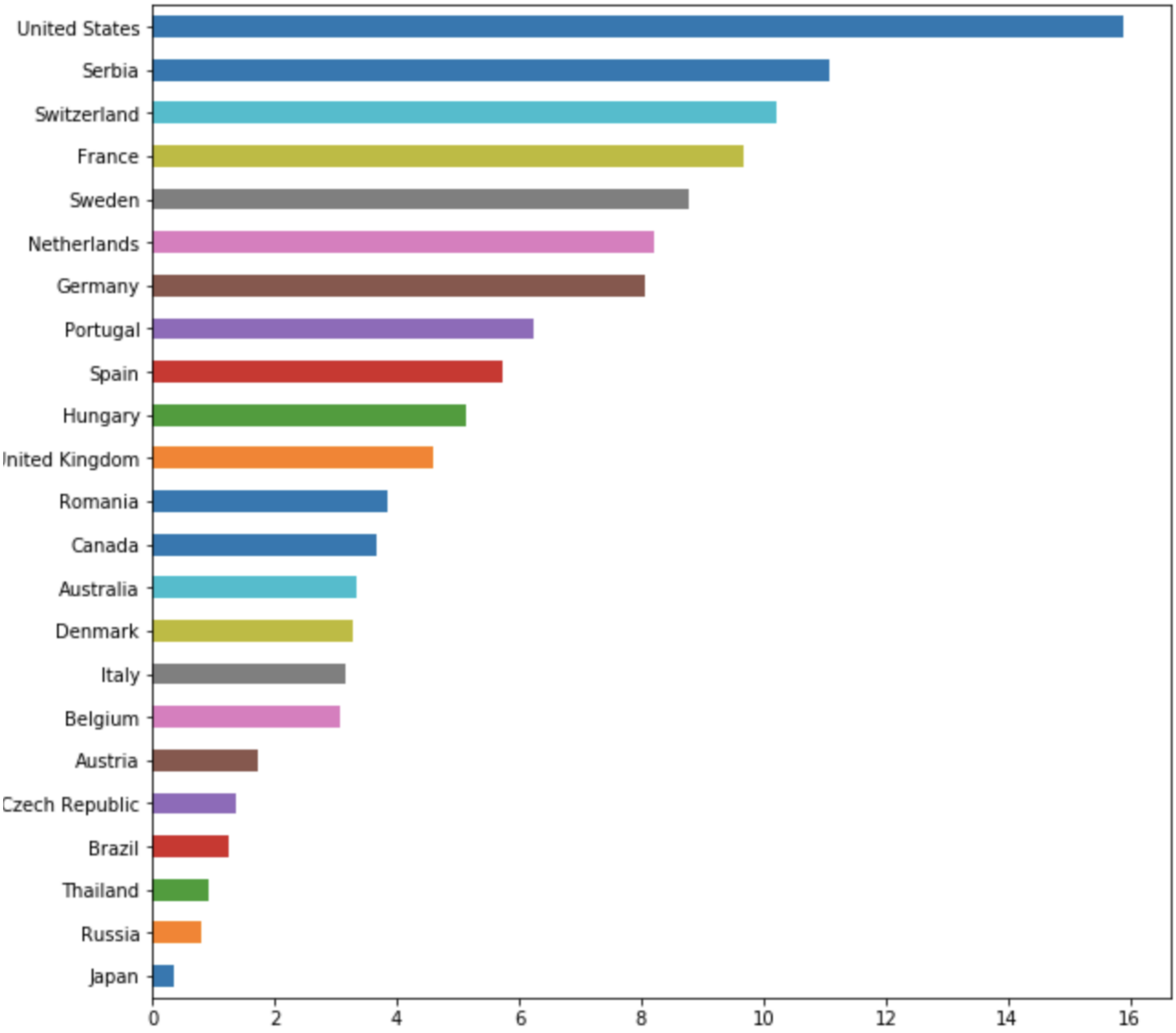


Figure 6: Top 5 countries with most sugar content [9]

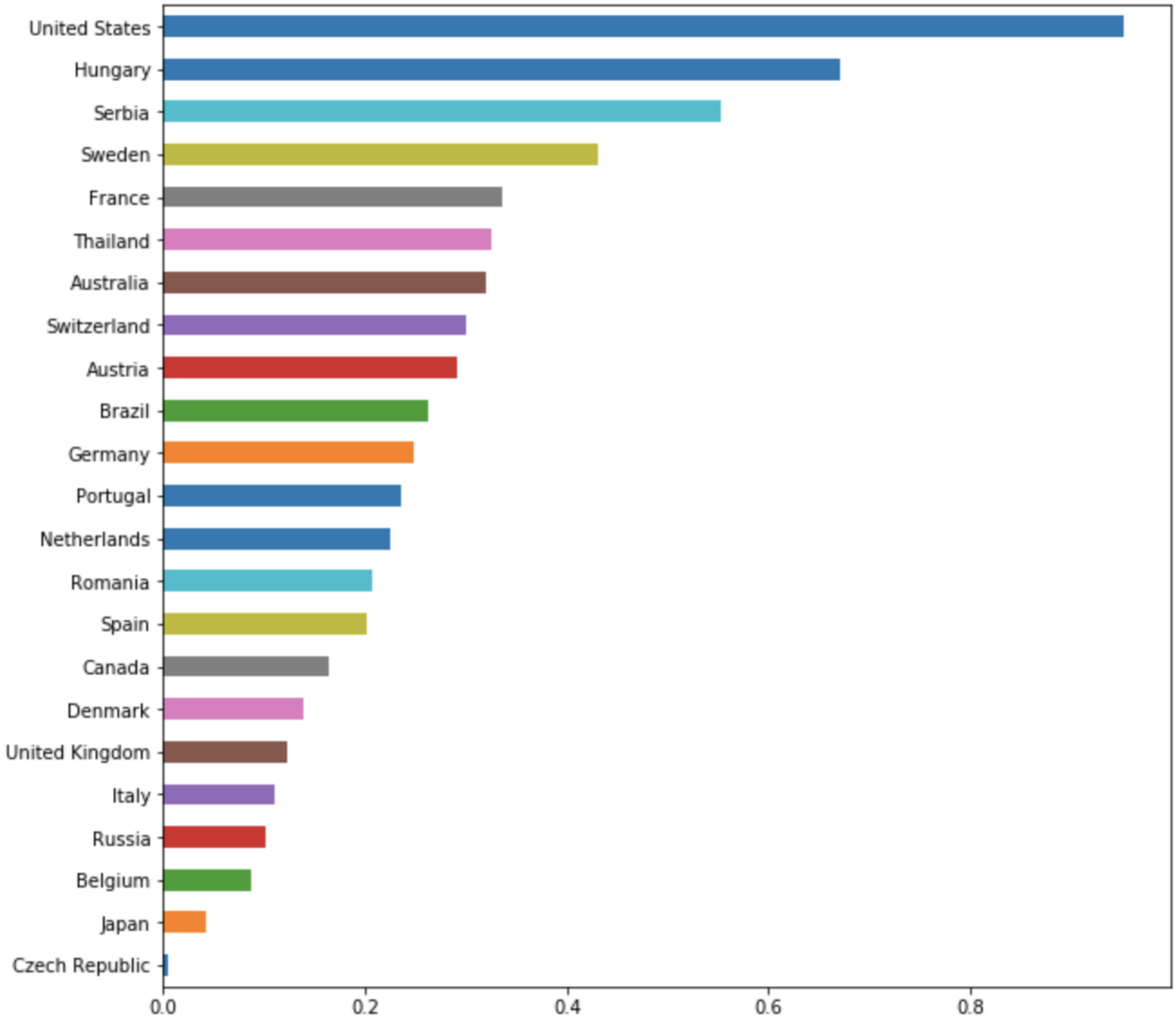
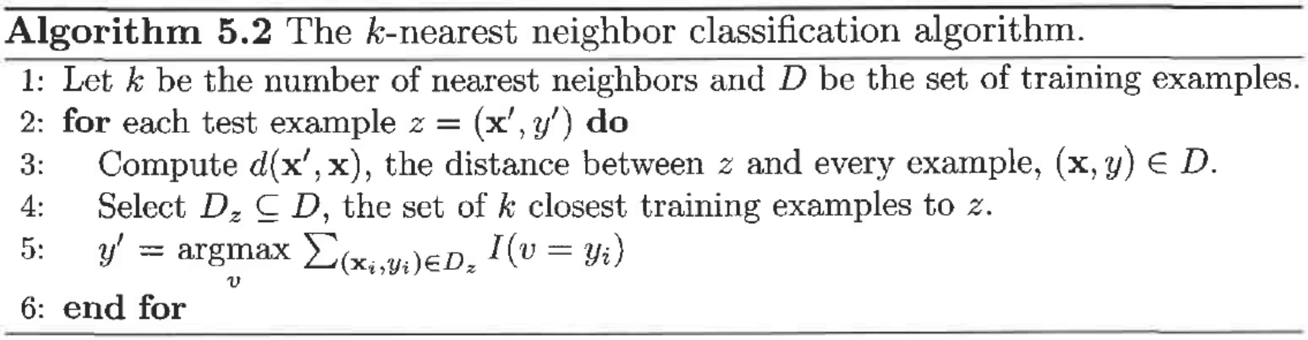


Figure 7: Top 5 countries with most sugar content [9]



Algorithm 5.2 The k-nearest neighbor classification algorithm.

```
Let $k$ be the number of nearest neighbors and $D$ be the set of training examples.
for each test example $z = (\mathbf{x}', y')$ do
    Compute $d(\mathbf{x}', \mathbf{x})$, the distance between $z$ and every example, $(\mathbf{x}, y) \in D$.
    Select $D_z \subseteq D$, the set of $k$ closest training examples to $z$.
    $y' = \underset{v}{\operatorname{argmax}} \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v = y_i)$
end for
```

Figure 8: K nearest neighbors algorithm[8]

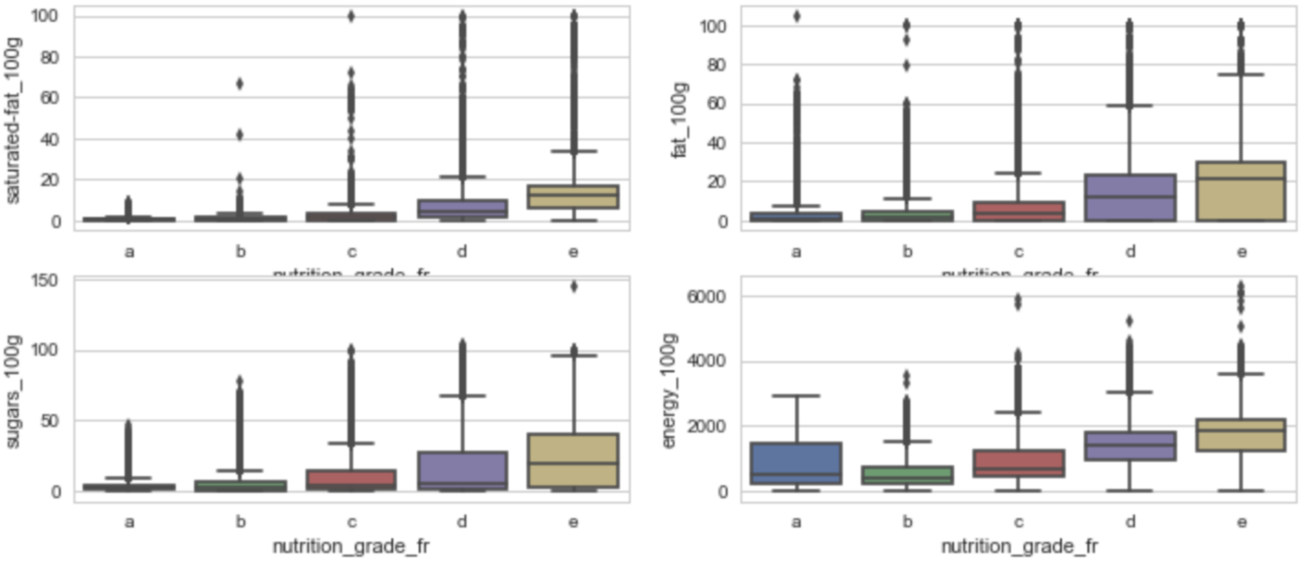
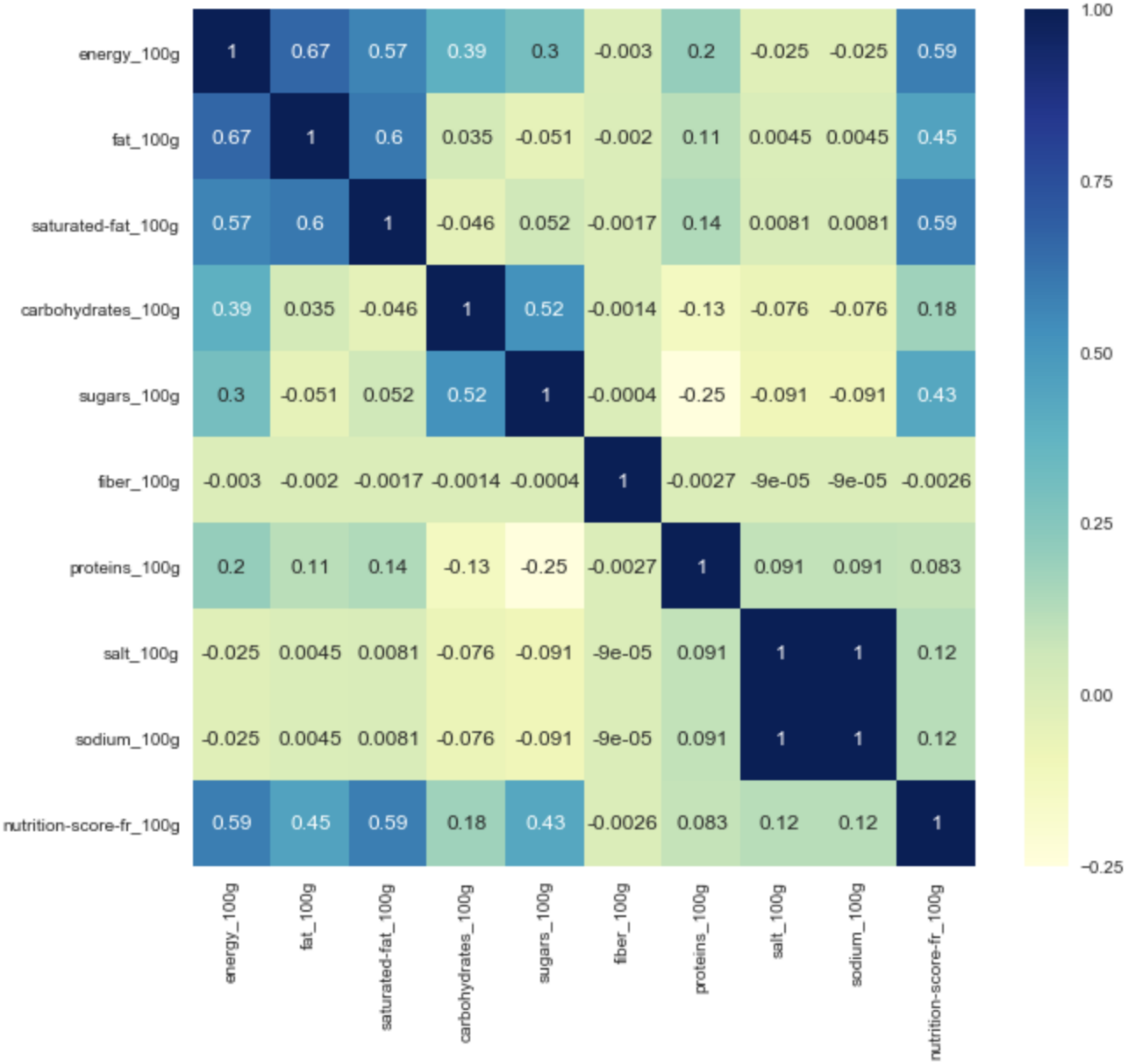


Figure 9: Bi-variate box plots [9]



	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g	sodium_100g	nutrition-score-fr_100g
energy_100g	1	0.67	0.57	0.39	0.3	-0.003	0.2	-0.025	-0.025	0.59
fat_100g	0.67	1	0.6	0.035	-0.051	-0.002	0.11	0.0045	0.0045	0.45
saturated-fat_100g	0.57	0.6	1	-0.046	0.052	-0.0017	0.14	0.0081	0.0081	0.59
carbohydrates_100g	0.39	0.035	-0.046	1	0.52	-0.0014	-0.13	-0.076	-0.076	0.18
sugars_100g	0.3	-0.051	0.052	0.52	1	-0.0004	-0.25	-0.091	-0.091	0.43
fiber_100g	-0.003	-0.002	-0.0017	-0.0014	-0.0004	1	-0.0027	-9e-05	-9e-05	-0.0026
proteins_100g	0.2	0.11	0.14	-0.13	-0.25	-0.0027	1	0.091	0.091	0.083
salt_100g	-0.025	0.0045	0.0081	-0.076	-0.091	-9e-05	0.091	1	1	0.12
sodium_100g	-0.025	0.0045	0.0081	-0.076	-0.091	-9e-05	0.091	1	1	0.12
nutrition-score-fr_100g	0.59	0.45	0.59	0.18	0.43	-0.0026	0.083	0.12	0.12	1

Figure 10: Correlation Plot [9]

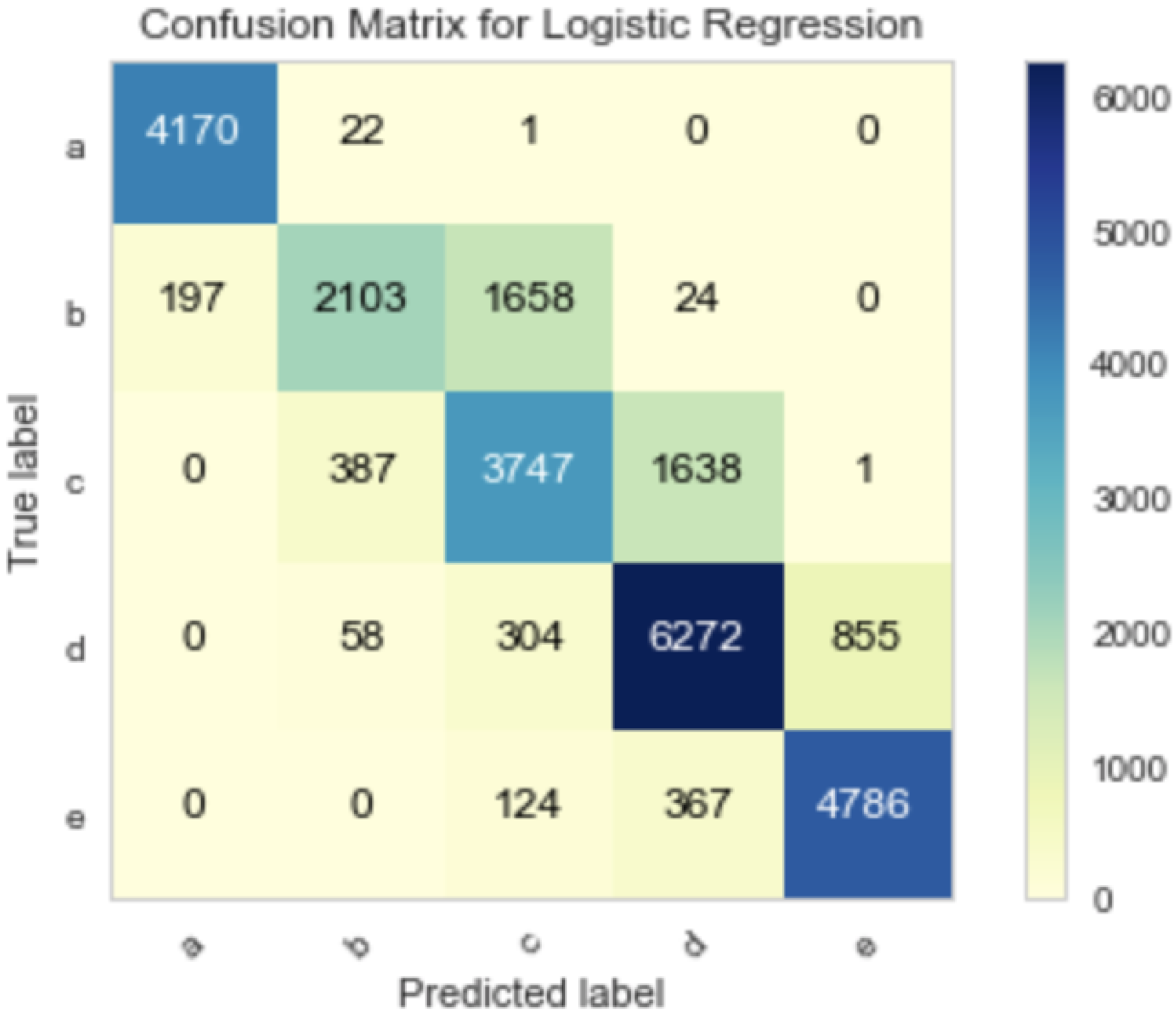


Figure 11: Confusion matrix for Logistic Regression [9]

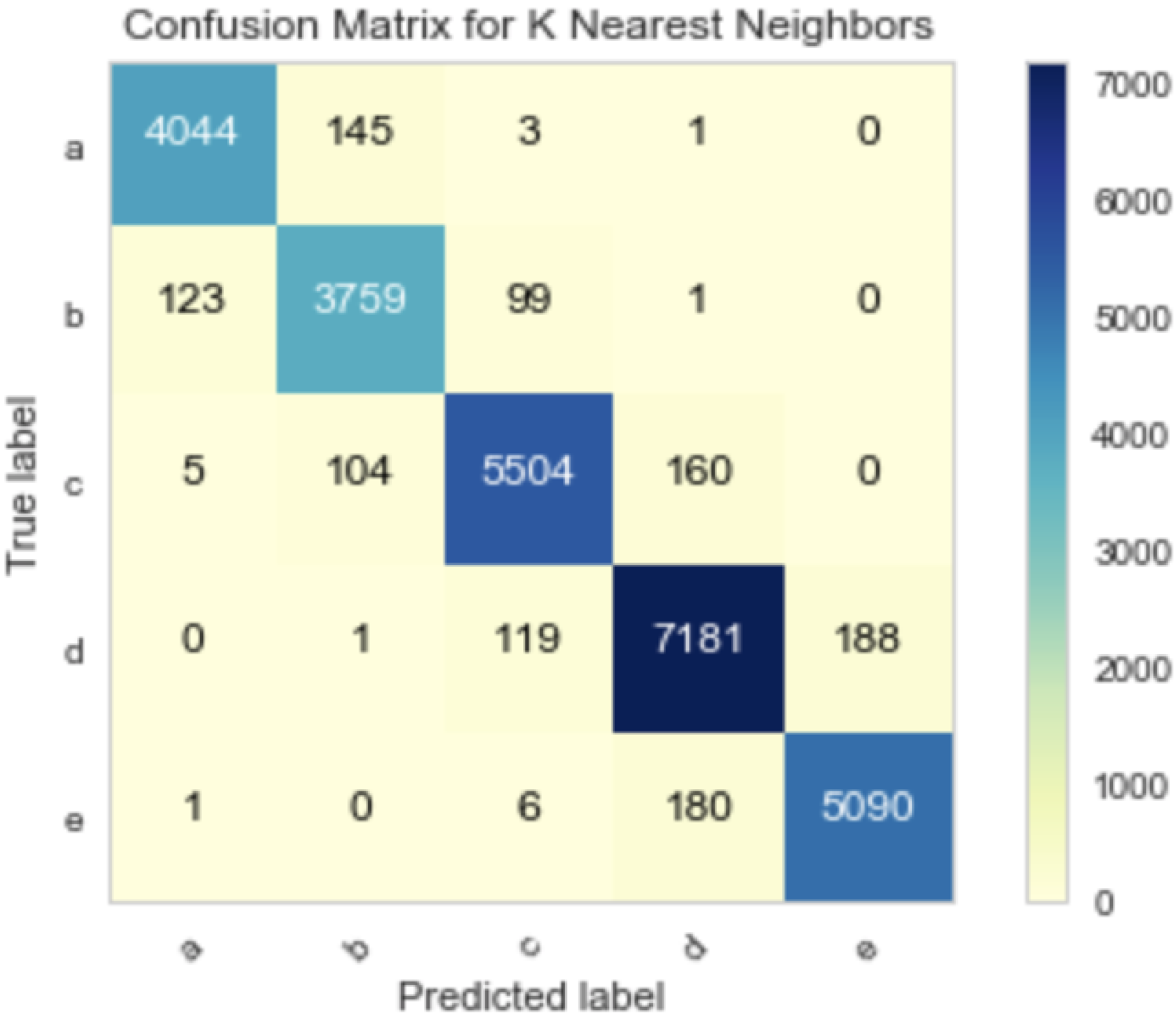


Figure 12: Confusion matrix for K Nearest Neighbors [9]

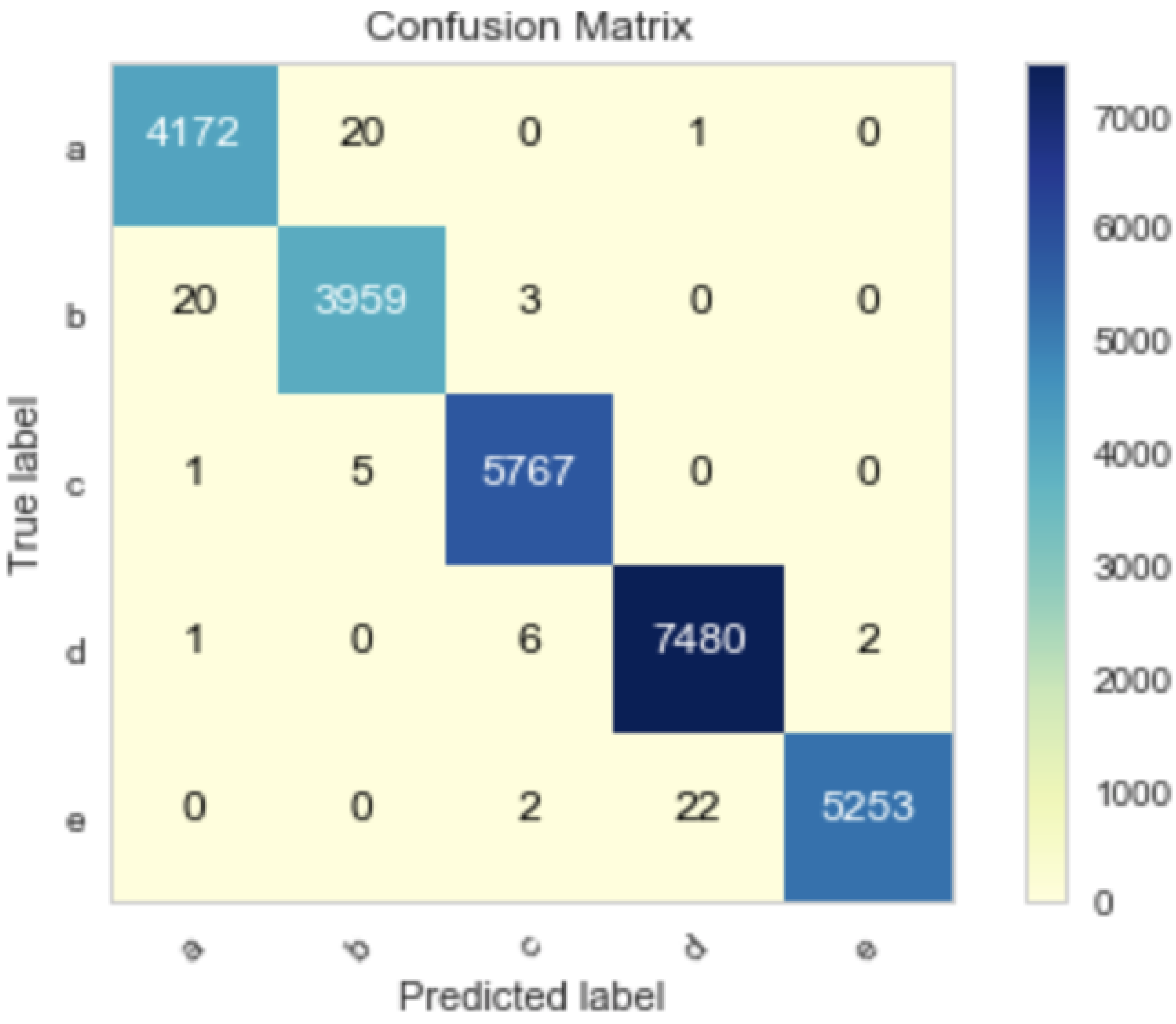


Figure 13: Confusion matrix for Random Forest [9]