Forecasting Flowering and Maturity Times of Barley Using Six Machine Learning Algorithms

Mingyuan Cheng and Mingchu Zhang

School of Natural Resources and Extension, University of Alaska Fairbanks, Fairbanks 99775-7200, Alaska, USA

Abstract: Interior Alaska has a short growing season of 110 d. The knowledge of timings of crop flowering and maturity will provide the information for the agricultural decision making. In this study, six machine learning algorithms, namely Linear Discriminant Analysis (LDA), Support Vector Machines (SVMs), $k$-nearest neighbor ($k$NN), Naïve Bayes (NB), Recursive Partitioning and Regression Trees (RPART), and Random Forest (RF), were selected to forecast the timings of barley flowering and maturity based on the Alaska Crop Datasets and climate data from 1991 to 2016 in Fairbanks, Alaska. Among 32 models fit to forecast flowering time, two from LDA, 12 from SVMs, four from NB, three from RF outperformed models from other algorithms with the highest accuracy. Models from $k$NN performed worst to forecast flowering time. Among 32 models fit to forecast maturity time, two models from LDA outperformed the models from other algorithms. Models from $k$NN and RPART performed worst to forecast maturity time. Models from machine learning methods also provided a variable importance explanation. In this study, four out of six algorithms gave the same variable importance order. Sowing date was the most important variable to forecast flowering but less important variable to forecast maturity. The daily maximum temperature may be more important than daily minimum temperature to fit flowering models while daily minimum temperature may be more important than daily maximum temperature to fit maturity models. The results indicate that models from machine learning provide a promising technique in forecasting the timings of flowering and maturity of barley.

Key words: Machine learning, flowering and maturity, Linear Discriminant Analysis, Support Vector Machines, $k$-nearest neighbor, Naïve Bayes, recursive partitioning, regression trees, Random Forest.

1. Introduction

Spring barley (Hordeum vulgare L.) is the most important grain crop grown in Alaska, 5,000 acres in 2016 according to annual U.S. Department of Agriculture statistics data [1]. The weather conditions for barley growth in the sub-arctic Alaska are low spring precipitation, cool summer temperature and cool-damp autumns. Barley has well adapted to the long day photoperiods (20+ hours of daylight in June) and a short frost-free growing season of 100-118 d. However, the short, cool growing season is thought as a barrier to barley growth in Alaska [2]. For example, the occurrence of frost in late spring during flowering can cause sterility, while light frost before plant reaches full maturity stops any further grain development [2]. Therefore, in Alaska, the knowledge of time of flowering and maturity of barley will aid farmers and scientists to know if the occurrence of flowering and maturity occurs during abiotic (such as cold, frost, heat and drought) and biotic stresses (including fungi, bacteria, viruses, nematodes, or insects).

Times to flowering and maturity are two fundamental and important crop traits to predict the crop yield. Understanding and predicting the flowering and maturity times and the important climatic factors affecting flowering and maturity times are crucial for optimizing crop management practices such as herbicide, fertilizer and irrigation applications. Genetic characteristics, agricultural management practices and environmental factor control times to flowering and maturity. Flowering time and maturity time are determined by a set of genes belonging to the
circadian clock, autonomous, ages and gibberellin pathways [3]. The agricultural management practices such as nitrogen deficiency and late sowing date also affect flowering time and maturity time [4, 5]. The environmental factors such as temperature, photoperiod and vernalization affect times to flowering and maturity [6]. Vernalization, a continuous chilling treatment, is required for the winter wheat and barley but not for spring wheat and barley [7]. Photoperiod is needed for the crops which are sensitive to the day and/or night length [7]. Temperature or growing degree days determine all developmental stages of crops [8]. Most crop growth models use temperature, photoperiod and vernalization to predict times of flowering and maturity [6, 9-14]. Only few studies use the temperature to predict the phonological stages of crops [14, 15]. In this study, temperature and sowing date were chosen to predict the timings of flowering and maturity of spring barley in interior Alaska.

Arctic is more sensitive to the climate changes than other places and the rate of temperature rise across the Arctic are twice the global average in past decades [16]. Alaska, a state ranged from subarctic to arctic regions, has experienced significant increases in temperature over much of the last century because of its geographic position [17]. The rising temperature causes the thawing of permafrost, expansion of tundra vegetation and land cover, greater growth of trees and expanding the growing season [18]. The expansion of the growing season may shift the phenological development stages of crops. Therefore, knowing flowering and maturity times of crops is of importance.

Many models are fit to forecast the flowering and maturity times. These models include simple linear models to complicated crop simulation models [6, 12, 19]. These models are either traditional statistics or mathematic models. With the increase of computer powers, machine learning algorithms provide a new approach to analyze data and give the prediction of the unknown data. Machine learning is a subfield of artificial intelligence which uses computer to learn and act like human being with the help of algorithms and data. A machine learning algorithm learns different properties of the data and discovers patterns of the data that can be late used to analyze new data in future. Machine learning is one of the fastest growing fields in the past couple of decades and is widely used in many areas. The applications of machine learning in agriculture include wheat yield prediction using artificial neural networks [20], rice development stage prediction and yield prediction using Support Vector Machines (SVMs) [21], soybean productive modeling using Decision Tree [22], rice Bakanae disease detection using SVM [23], wheat yellow rust detection using Reflectance Measurements and Neural Networks [24], wheat stripe rust prediction using SVM [25], weed detection using Artificial Neural Networks and Counter Propagation [26], coffees classification using Linear Discriminant Analysis (LDA) [27], crop classification using Random Forest (RF) [28], crop species recognition using Convolutional Neural Network [29], soil condition prediction using Least-Squares SVMs and Cubist [30], agricultural land soil classification using Naïve Bayes (NB) [31] and soil temperature estimation using Self-Adaptive Evolutionary Extreme Learning Machine [32]. Most of these applications use only one algorithm and few studies use multiple algorithms and compare their performance [33, 34].

In this study, six machine learning algorithms, namely LDA, SVMs, k-nearest neighbors (kNN), NB, Recursive Partitioning and Regression Trees (RPART) and RF, were selected to forecast flowering time and maturity time because these algorithms have been used in agriculture [27, 28, 31, 34, 35] and four of them (SVM, kNN, NB and RPART) are also on top 10 list of machine learning algorithms [36]. The aims of this study were to use the Alaska Crops Datasets from 1991 to 2016 and daily temperature data (1) to compare the performance of six machine learning
algorithms to forecast the flowering time and maturity time, (2) to explain the important variables related to the flowering and maturity.

2. Materials and Methods

2.1 Crop Information

Barley, cultivar Otal, was selected to forecast the timings of flowering and maturity. Sowing, flowering, and maturity dates were from Alaska Crops Databases from 1991 to 2016 in Fairbanks, Alaska and the dates were transformed to day of the year (DOY) as a continuous variable. The dates of flowering and maturity of Otal were equally divided into three groups, Early, Normal and Late. For flowering, Early group was the date flowering earlier than June 26, Normal group between June 26 and July 4, and Late group later than July 4. For maturity, Early group was the date maturing earlier than July 23, Normal group between July 23 and August 2 and Late group later than August 2.

2.2 Climate Data

The daily average, maximum and minimum temperatures were downloaded from Iowa Environmental Mesonet of Iowa State University [37]. Based on the data of flowering and maturity timings from 1991 to 2006, the minimum days from sowing to flowering and maturity were 35 d and 56 d, respectively. Therefore, the 30-day daily temperatures from sowing were selected to forest the timing of flowering and the 50-day daily temperature from sowing to forecast the timing of the maturity (Table 1). In order to test the importance of sowing in flowering and maturity, the date of sowing was included in some datasets. Four datasets were applied to fit and forecast flowering and four other datasets were applied to fit and forecast maturity (Table 1). Datasets to forecast flowering included 30-day average daily temperature with sowing date (F1 in Table 1) and without sowing date (F2 in Table 1), and 30-day daily minimum and maximum temperatures with sowing data (F3 in Table 1) and without sowing date (F4 in Table 1). Datasets to forecast maturity included 50-day daily average temperature with sowing date (M1 in Table 1) and without sowing date (M2 in Table 1), and 50-day minimum and maximum temperatures with sowing date (M3 in Table 1) and without sowing date (M4 in Table 1).

2.3 Algorithms

2.3.1 LDA

LDA is a very common technique for dimensionality reduction problems in machine learning originally developed in 1936 [38-40]. LDA technique searches for a lower-dimensional space which simply maximizes the separation distance between classes. There are three steps needed to perform the LDA [40]. The first step is to calculate distance between the means of different classes, which is called the between-class variance. The between-class variance of the $i$th class represents the distance between the mean of the $i$th class ($\mu_i$) and the

| Table 1  Dataset used in this study. |
|-----------------|------------------|
| Datasets | Crop information | Climate information |
| F1  | Sowing date and flowering date | 30-day daily average temperature from sowing |
| F2  | Flowering date | 30-day daily average temperature from sowing |
| F3  | Sowing date and flowering date | 30-day daily max and min temperatures from sowing |
| F4  | Flowering date | 30-day daily max and min temperatures from sowing |
| M1  | Sowing date and maturity date | 50-day daily average temperature from sowing |
| M2  | Maturity date | 50-day daily average temperature from sowing |
| M3  | Sowing date and maturity date | 50-day daily max and min temperatures from sowing |
| M4  | Maturity date | 50-day daily max and min temperatures from sowing |
based on their similarity with neighbors, and classification algorithms \[36\]. It classifies data sets maximum temperature). Late;

The distance between two points is calculated as the following:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2} \quad (4)$$

To predict an unknown object, the distance of this object to the known objects (from training set) is computed, its \(k\)NNs are identified, and the class categories of these nearest neighbors are then used to determine the class category of the object based on simple majority of these \(k\)NNs. In this study, \(k\) equals 5 or 7 depending on datasets (Table 2).

2.3.3 RPART

RPART is a type of supervised learning algorithm used in both classification and regression problems \[41, 42\]. In this technique, the population or sample is split into two or more homogeneous sets (or sub-populations) based on most significant splitter/differentiator in input variables. Gini index is one of the algorithms to determine the most significant variable for the decision tree split and it is used in this study. Gini index is a metric to measure how often a randomly chosen element would be correctly identified. It means an attribute with lower Gini index should be preferred. For any dataset \(X\), consisting of elements belonging to one of \(k\) classes, its Gini index can be calculated using the following formula:

$$\text{Gini} (P_1, \ldots, P_k) = 1 - \sum_{j=1}^{k} P_j^2 \quad (5)$$

where \(P_j\) is the relative frequency of class \(j\) from the set \(X\), belonging to the \(j\)th class \[43\]. Tests are done based on existing data, and when new data are being added, it can be classified to the corresponding group. In this paper \(j = 1, 2, 3\) represents three different groups. Dataset \(X\) represents the vector of explanatory variables (sowing date, daily average temperature, and daily minimum and maximum temperature).

2.3.4 RF

RF was devised by Breiman in the early 2000s \[44\] and is a type of ensemble machine learning algorithms that combine several machine learning techniques into one predictive model in order to decrease variance, bias or improve predictions. An RF uses many decision tree models which were fitted by bootstrap sampling the training data. Two steps are involved with random selection when forming the trees in the
Table 2  The optimal parameter values used in models to forecast the flowering and maturity time.

| Algorithms               | Data | Parameters                                      |
|--------------------------|------|------------------------------------------------|
|                          |      |                                                |
| **Linear Discriminant    |      |                                                |
| Analysis (LDA)           |      |                                                |
| F1                       |      | Not available                                  |
| F2                       |      | Not available                                  |
| F3                       |      | Not available                                  |
| F4                       |      | Not available                                  |
| M1                       |      | Not available                                  |
| M2                       |      | Not available                                  |
| M3                       |      | Not available                                  |
| M4                       |      | Not available                                  |
| **Support Vector Machine |      |                                                |
| (SVM) radial             |      |                                                |
| F1                       |      | Sigma = 0.01819, cost = 1                      |
| F2                       |      | Sigma = 0.01863, cost = 1                      |
| F3                       |      | Sigma = 0.00808, cost = 1                      |
| F4                       |      | Sigma = 0.00815, cost = 1                      |
| M1                       |      | Sigma = 0.01003, cost = 1                      |
| M2                       |      | Sigma = 0.01045, cost = 1                      |
| M3                       |      | Sigma = 0.00456, cost = 1                      |
| M4                       |      | Sigma = 0.00461, cost = 1                      |
| **SVM linear**           |      |                                                |
| F1                       |      | Cost = 1                                       |
| F2                       |      | Cost = 1                                       |
| F3                       |      | Cost = 1                                       |
| F4                       |      | Cost = 1                                       |
| M1                       |      | Cost = 1                                       |
| M2                       |      | Cost = 1                                       |
| M3                       |      | Cost = 1                                       |
| M4                       |      | Cost = 1                                       |
| **SVM polynomial**       |      |                                                |
| F1                       |      | Degree = 1, scale = 0.1, cost = 0.5           |
| F2                       |      | Degree = 1, scale = 0.1, cost = 0.5           |
| F3                       |      | Degree = 1, scale = 0.1, cost = 0.25          |
| F4                       |      | Degree = 1, scale = 0.1, cost = 0.25          |
| M1                       |      | Degree = 3, scale = 0.1, cost = 0.25          |
| M2                       |      | Degree = 3, scale = 0.1, cost = 0.25          |
| M3                       |      | Degree = 2, scale = 0.1, cost = 0.25          |
| M4                       |      | Degree = 2, scale = 0.1, cost = 0.25          |
| **k-nearest neighbor (KNN)** |  |                                                |
| F1                       |      | $k = 5$                                        |
| F2                       |      | $k = 7$                                        |
| F3                       |      | $k = 5$                                        |
| F4                       |      | $k = 5$                                        |
| M1                       |      | $k = 7$                                        |
| M2                       |      | $k = 7$                                        |
| M3                       |      | $k = 7$                                        |
| M4                       |      | $k = 5$                                        |
| **Naïve Bayes (NB)**     |      |                                                |
| F1                       |      | $fL = 0$, usekernel = FALSE and adjust = 1     |
| F2                       |      | $fL = 0$, usekernel = FALSE and adjust = 1     |
| F3                       |      | $fL = 0$, usekernel = FALSE and adjust = 1     |
| F4                       |      | $fL = 0$, usekernel = FALSE and adjust = 1     |
| M1                       |      | $fL = 0$, usekernel = TRUE and adjust = 1      |
| M2                       |      | $fL = 0$, usekernel = TRUE and adjust = 1      |
| M3                       |      | $fL = 0$, usekernel = TRUE and adjust = 1      |
| M4                       |      | $fL = 0$, usekernel = TRUE and adjust = 1      |
(Table 2 continued)

| Algorithms                  | Data | Parameters |
|-----------------------------|------|------------|
| Recursive Partitioning and  |      |            |
| Regression Trees (RPART)    |      |            |
| F1                          | CP   | = 0        |
| F2                          | CP   | = 0        |
| F3                          | CP   | = 0        |
| F4                          | CP   | = 0        |
| M1                          | CP   | = 0        |
| M2                          | CP   | = 0        |
| M3                          | CP   | = 0        |
| M4                          | CP   | = 0        |
| Random Forest (RF)          |      |            |
| F1                          | mtry | = 17       |
| F2                          | mtry | = 16       |
| F3                          | mtry | = 2        |
| F4                          | mtry | = 2        |
| M1                          | mtry | = 2        |
| M2                          | mtry | = 2        |
| M3                          | mtry | = 2        |
| M4                          | mtry | = 2        |

CP = complexity parameter is used to control the size of the decision tree and to select the optimal tree size; fL: factor for Laplace correction, default factor is 0 (no correction); mtry = number of variables randomly sampled as candidates at each split.

2.3.5 NB

NB, a classification technique, is based on Bayes’ theorem with the “naïve” assumptions that each input variable is independent [45]. Given a class variable \( Y (y_1, \ldots, y_j) \) and a dependent feature vector \( X (x_1, \ldots, x_n) \), Bayes’ theorem is as following:

\[
P(Y|X) = \frac{P(Y|X)p(x|Y)}{P(X)}
\]

where, \( P(Y|X) \) is the posterior probability of class given predictor; \( P(Y) \) is the prior probability of class; \( P(X|Y) \) is the likelihood which is the probability of predictor given class; \( P(X) \) is the prior probability of predictor.

The NB independent assumption is:

\[
P(X|Y) = P(x_1|y_1)P(x_2|y_2) \times \ldots \times P(x_n|y_j)
\]

\[
P(Y) \propto P(Y)\prod_{i=1}^{m} P(x_i|y_j)
\]

\[
y_{\text{prediction}} = \arg\max P(Y)\prod_{i=1}^{m} P(x_i|y_j)
\]

The posterior probability for each class is calculated and the class with the highest posterior probability is the outcome of prediction. In this study, \( Y_j \) represents classes \((j = 1, 2, 3)\) and \( X_i \) represents temperature and sowing date.

2.3.6 SVMs

The SVMs were presented by Cortes and Vapnik in the middle 1990s [46]. The SVMs are a set of supervised leaning methods used for classification, regression and outliers detection [46, 47]. In this paper, support vector classification was used. Suppose there is a set of \( m \) data points \( \{x_i, y_i\}_{i=1}^{m} \) with input data \( x_i \in \mathbb{R}^p \). The total number of data patterns and the output \( y_i \in \mathbb{R} \). In classification, a linear classifier is based on a linear discriminant function of the form.

\[
f(x) = w^T x + b
\]

where \( w^T x = \sum_i w_i x_i \) is known as the weight vector, and \( b \) is called the bias.

SVMs linearly separate the different classes of data by a hyperplane.
\{x: f(x) = w^T x + b = 0\} \quad (10)

To obtain \(w\) and \(b\),

\[
\text{minimize} \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} \xi_i \quad (11)
\]

subject to: \(y_i(w^T x + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \ldots, m\) \quad (12)

where \(\xi_i\) is the slack variable and \(C\) is the penalty coefficient which should be set artificially in practice.

The decision function for SVMs:

\[
f(x) = \text{sign}(w^T x + b) \quad (13)
\]

In many applications a non-linear classifier provides better accuracy and the kernel tricks are used. \(K(x_i, x_j)\) is defined as the kernel function which returns the inner product between two points in a suitable space. The commonly used kernel methods include the following:

The linear kernel:

\[
K(x_i, x_j) = x_i^T x_j \quad (14)
\]

The polynomial kernel:

\[
K(x_i, x_j) = \gamma(x_i^T x_j + c)^d \quad (15)
\]

where \(d\) is a tunable parameter;

The radial basis function (RBF):

\[
K(x_i, x_j) = \exp(-r \|x_i - x_j\|^2/2\sigma^2) \quad (16)
\]

with the parameter \(r > 0\). In this study, \(Y_j\) represents classes \((j = 1, 2, 3)\) and \(X_i\) represents temperature and sowing date. Among these kernel functions, RBF is used as default function in some of R packages [48].

2.4 Statistical Analysis

The classification models were constructed using the “caret” package [49] in R version 3.5.1 [50]. The machine learning algorithms used in this study were LDA, SVMs, kNN, NB, RPART and RF. Each dataset was randomly split into two parts, two-third for training sets and one-third for validation sets. Before modeling, the values of predictor variables in the datasets were pre-processed using “center” and “scale” method in “preprocess” function. The method “center” was to subtract the mean of the predictors’ data from the predictor values while method “scale” was to divide by the standard deviation. The “train” function from “caret” was used to select the values of model tuning parameters and estimate model performance using resampling method (10-fold cross-validation with 10 repeats). The optimal parameters used in each model were shown in Table 2. The metric, “accuracy”, was to select the optimal model. The final optimal model from each algorithm was used to predict the timing of flowering and maturity. The importance of each variable was evaluated using function “varImp” in the “caret” package of R [49].

2.5 Performance of Models Evaluation

Accuracy and Kappa were used to evaluate the models fitted from different algorithms.

\[
\text{Accuracy} = \frac{\text{Sum of number from different categories which were correctly predicted by the models/Total observations}}{1} \quad (17)
\]

The Kappa statistic is a metric that compares an observed accuracy with an expected accuracy.

\[
\text{Kappa} = \frac{(\text{Observed accuracy} - \text{Expected accuracy})}{(1 - \text{Expected accuracy})} \quad (18)
\]

3. Results

3.1 Performance of Different Algorithms to Forecast Flowering and Maturity Times

3.1.1 LDA Models

The performance of LDA depended on the data used and whether to forecast flowering time or maturity time. To forecast flowering time, the accuracy was 96%-100% and Kappa value was 0.91-1 for the training sets while the accuracy 96%-100% and Kappa 0.89-1 for validation sets (Table 3). However, when forecasting the maturity, the accuracy was 92% and Kappa value was 0.87 for the training sets while the accuracy was 76%-92% and Kappa value was 0.63-0.87 for validation sets (Table 4).

3.1.2 SVM Functions

Three SVM kernel functions (radial, linear and polynomial) were used to fit the forecasting models. The performance of these functions varied with data and whether to forecast flowering time or the maturity time. When SVMs were used to forecast flowering time, SVM linear and polynomial showed
a higher accuracy and Kappa values than SVM radial function for training sets. The accuracy was 100% and the Kappa value 1 for both SVM linear and polynomial functions for training sets (Table 3). However, the models from all three functions showed the same accuracy (100%) and Kappa value (1) from validation sets (Table 3). When SVMs were used to forecast maturity time, the accuracy and Kappa values were lower than those to forecast flowering time and also varied with functions. SVM linear performed better than the other two functions in forecasting maturity (Table 4). The accuracy ranged from 90% to 94% and Kappa value ranged from 0.84 to 0.91 for training sets while the accuracy 60% to 80% and Kappa value 0.37 to 0.7 for validation sets (Table 4).

Table 3  Performance of different algorithms to forecast the flowering time.

| Algorithms | Data | Training sets | Validation sets |
|------------|------|---------------|-----------------|
|            |      | Accuracy (95% CI) | Kappa | Accuracy (95% CI) | Kappa |
| LDA        | F1   | 100 (93 100)  | 1     | 96 (80 99)     | 0.89  |
|            | F2   | 100 (93 100)  | 1     | 96 (80 99)     | 0.89  |
|            | F3   | 96 (86 99)    | 0.91  | 100 (86 100)   | 1     |
|            | F4   | 100 (93 100)  | 1     | 100 (86 100)   | 1     |
| SVM radial | F1   | 98 (89 100)  | 0.96  | 100 (86 100)   | 1     |
|            | F2   | 96 (86 99)    | 0.91  | 100 (86 100)   | 1     |
|            | F3   | 96 (86 99)    | 0.91  | 100 (86 100)   | 1     |
|            | F4   | 96 (86 99)    | 0.91  | 100 (86 100)   | 1     |
| SVM linear | F1   | 100 (93 100)  | 1     | 100 (86 100)   | 1     |
|            | F2   | 100 (93 100)  | 1     | 100 (86 100)   | 1     |
|            | F3   | 100 (93 100)  | 1     | 100 (86 100)   | 1     |
|            | F4   | 100 (93 100)  | 1     | 100 (86 100)   | 1     |
| SVM poly   | F1   | 100 (93 100)  | 1     | 100 (86 100)   | 1     |
|            | F2   | 100 (93 100)  | 1     | 100 (86 100)   | 1     |
|            | F3   | 100 (93 100)  | 1     | 100 (86 100)   | 1     |
|            | F4   | 98 (89 99)    | 1     | 100 (86 100)   | 1     |
| kNN        | F1   | 88 (76 96)    | 0.75  | 88 (69 98)     | 0.67  |
|            | F2   | 76 (62 87)    | 0.41  | 88 (69 98)     | 0.61  |
|            | F3   | 78 (64 88)    | 0.56  | 76 (55 91)     | 0.47  |
|            | F4   | 78 (64 88)    | 0.56  | 68 (47 85)     | 0.37  |
| NB         | F1   | 100 (93 100)  | 1     | 100 (86 100)   | 1     |
|            | F2   | 100 (93 100)  | 1     | 100 (86 100)   | 1     |
|            | F3   | 100 (93 100)  | 1     | 100 (86 100)   | 1     |
|            | F4   | 100 (93 100)  | 1     | 100 (86 100)   | 1     |
| RPART      | F1   | 86 (73 94)    | 0.72  | 88 (69 98)     | 0.70  |
|            | F2   | 86 (73 94)    | 0.72  | 88 (69 98)     | 0.70  |
|            | F3   | 92 (81 98)    | 0.84  | 92 (74 99)     | 0.80  |
|            | F4   | 92 (81 98)    | 0.84  | 92 (74 99)     | 0.80  |
| RF         | F1   | 100 (93 100)  | 1     | 100 (86 100)   | 1     |
|            | F2   | 100 (93 100)  | 1     | 96 (80 99)     | 0.89  |
|            | F3   | 100 (93 100)  | 1     | 100 (86 100)   | 1     |
|            | F4   | 100 (93 100)  | 1     | 100 (86 100)   | 1     |
### Table 4  Performance of different algorithms to forecast the maturity time.

| Algorithms | Data | Training sets | Validation sets |
|------------|------|---------------|-----------------|
|            |      | Accuracy (95% CI) | Kappa | Accuracy (95% CI) | Kappa |
|            |      | %                |       | %                |       |
| LDA        | M1   | 92 (81 98)       | 0.87  | 76 (55 91)       | 0.63  |
|            | M2   | 92 (81 98)       | 0.87  | 76 (55 91)       | 0.63  |
|            | M3   | 92 (81 98)       | 0.87  | 92 (74 99)       | 0.87  |
|            | M4   | 92 (81 98)       | 0.87  | 92 (74 99)       | 0.87  |
| SVM radial | M1   | 90 (78 97)       | 0.84  | 60 (39 79)       | 0.37  |
|            | M2   | 90 (78 97)       | 0.84  | 60 (39 79)       | 0.37  |
|            | M3   | 92 (81 98)       | 0.87  | 80 (59 93)       | 0.68  |
|            | M4   | 92 (81 98)       | 0.87  | 72 (51 88)       | 0.56  |
| SVM linear | M1   | 94 (84 99)       | 0.91  | 76 (55 91)       | 0.63  |
|            | M2   | 94 (84 99)       | 0.91  | 80 (59 93)       | 0.70  |
|            | M3   | 94 (84 99)       | 0.91  | 76 (55 91)       | 0.63  |
|            | M4   | 94 (84 99)       | 0.91  | 76 (55 91)       | 0.63  |
| SVM poly   | M1   | 94 (84 99)       | 0.91  | 68 (47 85)       | 0.51  |
|            | M2   | 94 (84 99)       | 0.91  | 68 (47 85)       | 0.51  |
|            | M3   | 94 (84 99)       | 0.91  | 68 (47 85)       | 0.51  |
|            | M4   | 94 (84 99)       | 0.91  | 68 (47 85)       | 0.51  |
| kNN        | M1   | 78 (64 88)       | 0.66  | 52 (31 72)       | 0.27  |
|            | M2   | 70 (55 82)       | 0.54  | 52 (31 72)       | 0.27  |
|            | M3   | 82 (69 91)       | 0.72  | 40 (21 61)       | 0.06  |
|            | M4   | 74 (60 85)       | 0.60  | 40 (21 61)       | 0.07  |
| NB         | M1   | 94 (84 99)       | 0.91  | 72 (51 88)       | 0.57  |
|            | M2   | 94 (84 99)       | 0.91  | 72 (51 88)       | 0.57  |
|            | M3   | 94 (84 99)       | 0.91  | 80 (59 93)       | 0.69  |
|            | M4   | 92 (81 98)       | 0.88  | 80 (59 93)       | 0.69  |
| RPART      | M1   | 78 (64 88)       | 0.66  | 60 (39 79)       | 0.40  |
|            | M2   | 78 (64 88)       | 0.66  | 60 (39 79)       | 0.40  |
|            | M3   | 82 (69 91)       | 0.72  | 44 (24 65)       | 0.11  |
|            | M4   | 82 (69 91)       | 0.72  | 44 (24 65)       | 0.11  |
| RF         | M1   | 94 (84 99)       | 0.91  | 72 (51 88)       | 0.57  |
|            | M2   | 94 (84 99)       | 0.91  | 72 (51 88)       | 0.57  |
|            | M3   | 94 (84 99)       | 0.91  | 72 (51 88)       | 0.57  |
|            | M4   | 94 (84 99)       | 0.91  | 72 (51 88)       | 0.57  |

### 3.1.3 The kNN Methods

The kNN performed worse than other algorithms and the accuracy and Kappa values were low for forecasting both flowering time and maturity time. The accuracy ranged from 76% to 88% and Kappa value ranged from 0.41 to 0.75 for training sets while the accuracy 68% to 88% and Kappa value 0.37 to 0.67 for validation sets when kNN was used to forecast flowering time (Table 3). The accuracy ranged from 70% to 82% and Kappa value ranged from 0.54 to 0.72 for training sets while the accuracy 40% to 52% and Kappa value 0.06 to 0.27 for validation sets when kNN was used to forecast maturity (Table 4).

### 3.1.4 NB Performance

The performance of NB models to forecast flowering time was the best in all algorithms, with the 100% of accuracy and 1 of Kappa value for both training sets and validation sets (Table 3). The performance of NB model to forecast maturity time...
was better than SVM polynomial function, \(k\)NN, RPART and RF models. The accuracy ranged from 92% to 94% for training sets and the accuracy 72% to 80% for validation sets while the Kappa value ranged from 0.88 to 0.91 for training sets and from 0.57 to 0.69 for validation sets (Table 4).

3.1.5 RPART

RPART models were consistent to forecast flowering time for both training sets and validation sets. The accuracy was 86%-92% and Kappa value was 0.72-0.84 for training sets while the accuracy 88%-92% and Kappa value 0.70-0.80 for the validation sets (Table 3). To forecast maturity time, the performance of RPART models was much better for training sets than for validation sets, with 78%-82% accuracy and 0.66-0.72 Kappa value for training sets but only 44%-60% accuracy and 0.11-0.40 Kappa value for validation (Table 4).

3.1.6 RF Algorithm

RF was one of the best algorithms to forecast flowering time. The accuracy was 100% and Kappa value 1 for training set while the accuracy 96%-100% and Kappa value 0.89-1 for validation sets (Table 3). In addition, the performance of RF to forecast maturity was consistent in all four datasets. The accuracy was also 94% and Kappa value was 0.91 for training sets while the accuracy 72% and Kappa value 0.57 for validation sets (Table 4).

3.2 Variable Importance to Fit Flowering and Maturity Models

Models from all algorithms showed that the most important variable for forecasting flowering time was sowing date (Figs. 1 and 2). The daily average temperatures of the first week and the last week of 30-day period were important for flowering (Fig. 1). Seven out of 10 daily maximum temperature variables were among the top 10 important variables to forecast flowering, indicating that the daily maximum temperature may be more important than daily minimum temperature for flowering (Fig. 2). The results of important variables for flowering from RPART and RF were different from those of all other models (Figs. 1 and 2).

Sowing was still one of important variables for maturity but it was not as important as for flowering (Figs. 3 and 4). The daily average temperature of the second week after sowing and the daily average temperature just before flowering (30 d after sowing) were among the top 10 important variables (Fig. 3). Seven out of 10 top 10 important variables to forecast maturity were daily minimum temperature, indicating daily minimum temperature may be more important than daily maximum temperature to fit maturity models (Fig. 4). Model results from RPART and RF also showed different orders of important variables from other models (Figs. 3 and 4).

4. Discussion

4.1 Performance Comparison among Algorithms

No single learning algorithm can uniformly outperform other algorithms over all datasets. In this study, six machine learning algorithms namely LDA, SVMs, NB, \(k\)NN, RPART and RF have been compared in terms of accuracy and Kappa to forecast three categories (Early, Normal and Late) timings of flowering and maturity. SVMs and NB and RF were the best classifiers to provide the overall 100% correct classification for forecasting flowering time while LDA outperformed other algorithms to forecast maturity time. \(k\)NN was the worst algorithm to forecast flowering time while \(k\)NN and RPART were the worst algorithm to forecast maturity time. The reason SVMs perform better may be that SVM tends to perform much better when dealing with multi-dimensions and continuous feature [51] and in this paper the daily temperatures used to forecast times of flowering and maturity are continuous features and multicollinearity is also present. The reason that \(k\)NN has the worst performance may attribute to too much noise in daily temperature and \(k\)NN is generally considered intolerant of noise [51].
Fig. 1  Relative importance of variables in different algorithms when using data F1 to forecast flowering: (a) Linear Discriminant Analysis (LDA), (b) Support Vector Machines (SVM) radial function, (c) SVM linear function, (d) SVM polynomial function, (e) k-nearest neighbor (kNN), (f) Naïve Bayes (NB), (g) Recursive Partitioning and Regression Trees (RPART) and (h) Random Forest (RF).

4.2 Performance of SVMs

In this study, three kernel functions (i.e., radial, linear and polynomial) were selected to forecast timings of flowering and maturity. The performances of these kernels varied with datasets, especially for flowering. Radial function and linear function outperformed polynomial function in flowering forecast. These results indicate that multiple functions should be tested when SVMs are used to fit and predict classification.

4.3 Overfitting Problems

Overfitting in machine learning refers to a model that does so well in training set but does poor in validation...
set [52]. It happens when machine learning is supervised learning—that is, learning from labeled training data and a model learns the detail and noise in the training data [53]. In this paper, models from all algorithms tended to overfitting. The worst one was models of $k$NN to forecast maturity time with only 40%-52% of accuracy in validation sets (Table 4). Cross-validation for training data, adding a regularization term (or penalization) to the evaluation function and balancing machine learning training data may be used to overcome overfitting problems [52, 54].

4.4 Differences between Machine Learning and Traditional Statistics in Input and Output

There are many differences between machine learning and traditional statistics. Two differences between machine learning and traditional statistics are input and output based on this study. In input, machine learning prefers to work with abundant data,
often not much pre-assumptions and data distributions while traditional statistics prefers to work with small samples, prior knowledge of pre-assumptions and data distributions. In this study, the variables to build the models through algorithms were continuous daily temperature. These variables are time series and the variables are multicollinearity. If the traditional statistics were used to fit these models (for examples fit a multiple linear model), the data would violate the assumption of independent variable. Overall prediction is not affected by multicollinearity but the interpretation of conclusions based on the regression coefficients, the standard errors or \( t \)-test may be misleading because of the potentially confounding effect of multicollinearity [55]. In output, machine learning generalizes the patterns from data through computer algorithms and makes predictions. However, the output of traditional statistics model is an equation

Fig. 3  Relative importance of variables in different algorithms when using data M1 to forecast maturity: (a) LDA, (b) SVM radial function, (c) SVM linear function, (d) SVM polynomial function, (e) \( k \)NN, (f) NB, (g) RPART and (h) RF.
with only several independent variables. In general, machine learning does well with high number of variables and high number of observations.

4.5 More Observations and Data Features

The accuracy was lower to forecast maturity than to flowering for all algorithms. The highest accuracy from validation sets was 92% for flowering time while 100% for flowering time, indicating that temperature alone may not be enough to forecast maturity time. In order to improve the accuracy of validation sets more observations (in this case more years’ observations) and more predictor variables such as rain fall, fertilizer and irrigation data may help improve the accuracy of forecasting.

Fig. 4 Relative importance of variables in different algorithms when using data M3 to forecast maturity: (a) LDA, (b) SVM radial function, (c) SVM linear function, (d) SVM polynomial function, (e) kNN, (f) NB, (g) RPART and (h) RF.
4.6 Sowing Date and Flowering and Maturity Time

Sowing date affects times of flowering and maturity. Time of flowering occurs earlier than the optimum time when a spring wheat variety is sown early [56]. In this study, variables importance reported from models indicates that sowing date is the most important variable to forecast flowering time and less important variable to forecast maturity time. The additional tests showed that sowing date was positively related with both flowering time and maturity time (Fig. 5). However, sowing date was highly correlated with flowering date with the correlation coefficient 0.89 and moderately correlated with timing of maturity with the correlation coefficient 0.55, indicating that sowing date affects more on flowering time than on maturity time and there are more other factors to affect the timing of maturity.

4.7 Significance of Forecasting Times of Flowering and Maturity in Climate Change Scenarios

Alaska has experienced a statewide average warming of about 1 °C from 1920 to 2012 [17]. The warming temperature causes expanding growing season and increasing shrub abundance in the Arctic [57-59]. The rising temperature may also cause expanding growing season of small grain crops such as barleys, oats and wheats although there are no reports about the changes of phenological development stages related to climatic changes so far. The models built in this study can be used to predict the timings of flowering and maturity based on the

Fig. 5  Linear relationship of sowing date with flowering time (a) and maturity time (b).
simulation of the future climate change files. The information from forecasting models for flowering and maturity will help the researchers to decide if the late maturing varieties are needed, if the new breeding systems are needed, and how the new late maturing varieties will cope with the climatic change of Alaska. These forecasting models will also help farmers to use the varieties to cope with the climatic change, and to adapt the new agricultural management such as time of irrigation, chemical applications, fertilizer applications and time to harvest.

4.8 Future Work

The algorithms used in this study can be used as classification and regression. Classification method selected here to test and compare whether the machine learning methods were suitable to forecast flowering time and maturity time. Regression models can provide the accurate dates of flowering and maturity and may be good for decision making. In the future, the regression models should be fitted to better understand times of flowering and maturity. Another work needed to do in the future is using the models to forecast the times of flowering and maturity based on the simulation of climate change.

5. Conclusions

Machine learning algorithms are suitable to forecast flowering time and maturity time. No single learning algorithm can uniformly outperform other algorithms over all datasets. Among six machine learning algorithms, LDA and SVMs performed best and the performances of other algorithms depended on data and targets (both flowering and maturity). kNN was the worst algorithm in this study. Sowing date was the most important variable for flowering time. Daily maximum temperatures may be important for flowering models while daily minimum temperature may be important for maturity models.

Acknowledgments

The research is funded by United States Department of Agriculture (USDA) Hatch Project ALK 19-04 and Multi-state Project ALK 15-03.

References

[1] United States Department of Agriculture, National Agricultural Statistics Service. 2017. “Alaska Agricultural Statistics 2017 Annual Bulletin.” Accessed December 3, 2019. https://www.nass.usda.gov/Statistics_by_State/Alaska/Publications/Annual_Statistical_Bulletin/2017/AKANNUAL2017.pdf.

[2] Sharratt, B. 1999. “Thermal Requirement for Barley Maturation and Leaf Development in Interior Alaska.” *Field Crop. Res.* 63: 179-84.

[3] Hill, C. B., and Li, C. 2016. “Genetic Architecture of Flowering Phenology in Cereals and Opportunities for Crop Improvement.” *Front. Plant Sci.* 7: 1-23.

[4] Ma, Q., Longnecker, N., and Dracup, M. 1997. “Nitrogen Deficiency Slows Leaf Development and Delays Flowering in Narrow-Leafed Lupin.” *Ann. Botany* 79 (4): 403-9.

[5] Nielsen, R. L., Thomison, P. R., Brown, G. A., Halter, A. L., Wells, J., and Wuethrich, K. L. 2002. “Delayed Planting Effects on Flowering and Grain Maturation of Dent Corn.” *Agron. J.* 94 (3): 549-58.

[6] Palosuo, T., Kersebaum, K. C., Angulo, C., Hlavinka, P., Moriondo, M., Olesen, J. E., Patil, R. H., Ruget, F., Rumbaur, C., Jakáč, J., Trnka, M., Bindi, M., Caldař, B., Ewert, F., Ferrise, R., Mirschel, W., Saylan, L., Šiška, B., and Rötter, R. 2011. “Simulation of Winter Wheat Yield and Its Variability in Different Climates of Europe: A Comparison of Eight Crop Growth Models.” *Eur. J. Agron.* 35 (3): 103-14.

[7] Kamran, A., Iqbal, M., and Spaner, D. 2014. “Flowering Time in Wheat (*Triticum aestivum* L.): A Key Factor for Globe Adaptability.” *Euphytica* 197 (1): 1-26. doi:10.1007/s10681-014-1075-7.

[8] Hatfield, J. L., and Prueger, J. H. 2015. “Temperature Extremes: Effect on Plant Growth and Development.” *Weather and Climate Extremes* 10: 4-10.

[9] Wang, E., and Engel, T. 1998. “Simulation of Phonological Development of Wheat Crops.” *Agr. Syst.* 58 (1): 1-24.

[10] Juskiw, P. E., Jame, Y.-W., and Kryzanowski, L. 2001. “Phenological Development of Spring Barley in a Short-Season Growing Area.” *Agron. J.* 93 (2): 370-9.
Forecasting Flowering and Maturity Times of Barley Using Six Machine Learning Algorithms

[11] McMaster, G. C., and Wilhelm, W. W. 2003. “Phenological Responses of Wheat and Barley to Water and Temperature: Improving Simulation Model.” J. Agr. Sci. 141: 129-47.

[12] Yin, X., Struik, P. C., Tang, J., Qi, C., and Liu, T. 2005. “Model Analysis of Flowering Phenology in Recombinant Inbred Lines of Barley.” J. Exp. Bot. 56 (413): 959-65.

[13] He, L., Asseng, S., Zhao, G., Wu, D., Yang, X., Zhuang, W., Jin, N., and Yu, Q. 2015. “Impacts of Recent Climate Warming, Cultivar Changes, and Crop Management on Winter Wheat Phenology across the Loess Plateau of China.” Agr. Forest Meteorol. 200: 135-43.

[14] Sharifi, H., Hjimans, R. J., Hill, J. E., and Linquist, B. A. 2017. “Using Stage-Dependent Temperature Parameters to Improve Phenological Model Prediction Accuracy in Rice Models.” Crop Sci. 57 (1): 444-53.

[15] Shaykewich, C. F. 1995. “An Appraisal of Cereal Crop Phenology Modelling.” Can. J. Plant Sci. 75 (2): 329-41.

[16] Melvin, A. M., Larsen, P., Boehlert, B., Neumann, J. E., Chinowsky, P., Espinet, X., Martinich, J., Baumann, M. S., Rennels, L., Bothner, A., Nicolsky, D. J., and Marchenko, S. S. 2017. “Climate Change Damages to Alaska Public Infrastructure and the Economics of Proactive Adaptation.” PNAS 114 (2): E122-31. doi:10.1073/pnas.1610561113.

[17] Bieniek, P. A., Walsh, J. E., Thoman, R. L., and Bhatt, U. S. 2014. “Using Climate Divisions to Analyze Variations and Trend in Alaska Temperature and Precipitation.” J. Climate 27 (8): 2800-18. doi:10.1175/JCLI-D-13-00342.1.

[18] Hinzman, L. D., Bettez, N. D., Bolton, W. R., Chapin, F. S., Dyurgerov, M. B., Fastie, C. L., Griffith, B., Hollister, R. D., Hope, A., Huntington, H. P., Jensen, A. M., Jia, G. J., Jorgenson, T., Kane, D. L., Klein, D. R., Kofinas, G., Lynch, A. H., Lloyd, A. H., McGuire, A. D., Nelson, F. E., Oechel, W. C., Osterkamp, T. E., Racine, C. H., Romanovsky, V. E., Stone, R. S., Stow, D. A., Sturm, M., Tweedie, C. E., Vourlikis, G. L., Walker, M. D., Walker, D. A., Webber, P. J., Welker, J. M., Winker, K. S., and Yoshikawa, K. 2005. “Evidence and Implications of Recent Climate Change in Northern Alaska and Other Arctic Regions.” Climatic Change 72 (3): 251-98. doi:10.1007/s10584-005-5352-2.

[19] Oteros, J., Garcia-Mozo, H., Botey, R., Mestre, A., and Galán, C. 2015. “Variations in Cereal Crop Phenology in Spain over the Last Twenty-Six Years (1986-2012).” Clim. Change 130 (4): 545-58. doi:10.1007/s10584-015-1363-9.

[20] Pantazi, X. E., Moshou, D., Alexandridis, T. K., Whetten, R. L., and Mouazen, A. M. 2016. “Wheat Yield Prediction Using Machine Learning and Advanced Sensing Techniques.” Comput. Electron. Agric. 121: 57-65.

[21] Su, Y., Xu, H., and Yan, L. 2017. “Support Vector Machine-Based Open Crop Model (SBOCM): Case of Rice Production in China.” Saudi J. Biol. Sci. 24 (3): 537-47.

[22] Veenadhari, S., Mishra, B., and Singh, C. D. 2011. “Soybean Productivity Modelling Using Decision Tree Algorithms.” Int. J. Comput. Appl. 27 (7): 11-5.

[23] Chung, C. L., Huang, K. J., Chen, S. Y., Lai, M. H., Chen, Y. C., and Kuo, Y. F. 2016. “Detecting Bakanae Disease in Rice Seedlings by Machine Vision.” Comput. Electron. Agric. 121: 404-11.

[24] Moshou, D., Bravo, C., West, J., Wahlen, S., McCartney, A., and Ramon, H. 2004. “Automatic Detection of ‘Yellow Rust’ in Wheat Using Reflectance Measurements and Neural Networks.” Comput. Electron. Agric. 44 (3): 173-88.

[25] Wang, H., and Ma, Z. 2011. “Prediction of Wheat Stripe Rust Based on Support Vector Machine.” In Proceeding of 2011 Seventh International Conference on Natural Computation, 378-82.

[26] Pantazi, X. E., Tamouridou, A. A., Alexandridis, T. K., Lagopodi, A. L., Kashefi, J., and Moshou, D. 2017. “Evaluation of Hierarchical Self-Organising Maps for Weed Mapping Using UAS Multispectral Imagery.” Comput. Electron. Agric. 139: 224-30.

[27] Souto, U. T. C. P., Pontes, M. J. C., Silva, E. C., Galvão, R. K. H., Araújo, M. C. U., Sanches, F. A. C., Cunha, F. A. S., and Oliveira, M. S. R. 2010. “UV-Vis Spectrometric Classification of Coffees by SPA-LDA.” Food Chem. 119 (1): 368-71. doi:10.1016/j.foodchem.2009.05.078.

[28] Ok, A. O., Akar, O., and Gungor, O. 2012. “Evaluation of Random Forest Method for Agricultural Crop Classification.” Eur. J. Remote Sens. 45 (1): 421-32. doi:10.5721/EuJRS20124535.

[29] Grinblat, G. L., Uzal, L. C., Larese, M. G., and Granitto, P. M. 2016. “Deep Learning for Plant Identification Using Vein Morphological Patterns.” Comput. Electron. Agric. 127: 418-24.

[30] Morellos, A., Pantazi, X. E., Moshou, D., Alexandridis, T., Whetten, R., Tziotzios, G., Wiebensohn, J., Bill, R., and Mouazen, A. M. 2016. “Machine Learning Based Prediction of Soil Total Nitrogen, Organic Carbon and Moisture Content by Using VIS-NIR Spectroscopy.” Biosyst. Eng. 152: 104-16.

[31] Bhargavi, P., and Jyothi, S. 2009. “Applying Naïve Bayes Data Mining Technique for Classification of Agricultural
Land Soils.” Int. J. Comput. Sci. Netw. 9 (9): 117-22.

[32] Nahvi, B., Habibi, J., Mohammadi, K., Shamshirband, S., and Al Razgan, O. S. 2016. “Using Self-Adaptive Evolutionary Algorithm to Improve the Performance of an Extreme Learning Machine for Estimating Soil Temperature.” Comput. Electron. Agric. 124: 150-60.

[33] Nitze, I., Schulthess, U., and Asche, H. 2012. “Comparison of Machine Learning Algorithms Random Forest, Artificial Neural Network and Support Vector Machine to Maximum Likelihood for Supervised Crop Type Classification.” In Proceedings of the 4th GEOFIA, 35-40.

[34] Liakos, K. G., Busato, P., Moshou, D., Pearson, S., and Bochtis, D. 2018. “Machine Learning in Agriculture: A Review.” Sensors 18 (8): 2674.

[35] Patel, H., and Patel, D. 2014. “A Brief Survey of Data Mining Techniques Applied to Agricultural Data.” Int. J. Comput. Appl. 95 (9): 6-8.

[36] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., and Steinberg, D. 2008. “Top 10 Algorithms in Data Mining.” Knowl Inf Syst. 14 (1): 1-37. doi:10.1007/s10115-007-0114-2.

[37] Iowa State University, Iowa Environmental Mesonet. 2018. “ASOS Network/ASOS-AWOS-METAR Data Download.” Accessed July 17, 2018. https://mesonet.agron.iastate.edu/request/download.phtml?network=AK_ASOS.

[38] Fisher, R. A. 1936. “The Use of Multiple Measurements in Taxonomic Problems.” Ann. Eugenic. 7: 179-88.

[39] Welling, M. 2000. Fisher Linear Discriminant Analysis. Technical Report, University of Toronto, Toronto, Canada. https://www.ics.uci.edu/~welling/teaching/273ASpring09/Fisher-LDA.pdf.

[40] Tharwat, A., Gaber, T., Ibrahim, A., and Hassanien, A. E. 2017. “Linear Discriminant Analysis: A Detailed Tutorial.” AI Communications 30 (2): 169-90. doi:10.3233/AIC-170729.

[41] Therneau, T. M., and Atkinson, E. J. 1997. An Introduction to Recursive Partitioning Using the RPART Routines. Technical Report, Department of Health Sciences Research, Mayo Clinic. https://r.789695.n4.nabble.com/attachment/3209029/0/ze d.pdf.

[42] Kingsford, C., and Salzberg, S. 2008. “What Are Decision Trees?” Nat Biotechnol. 26: 1011-3. doi:10.1038/nbt0908-1011.

[43] Rutkowski, L., Pietruczuk, L., Duda, P., and Jaworski, M. 2013. “Decision Trees for Mining Data Streams Based on the McDiarmid’s Bound.” IEEE Trans. Knowl. Data Eng. 25 (6): 1272-9.

[44] Breiman, L. 2001. “Random Forests.” Mach. Learn. 45 (1): 5-32.

[45] Zhang, H. 2004. “The Optimality of Naïve Bayes.” In Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, Miami Beach, FL, USA. http://www.cs.umd.edu/~hzhang/publications/FLAIRS04Zh angH.pdf.

[46] Cortes, C., and Vapnik, V. N. 1995. “Support-Vector Networks.” Mach. Learn. 20 (3): 273-97. doi:10.1023/A:10933404324.

[47] Karatzoglou, A., Meyer, D., and Hornik, K. 2006. “Support Vector Machine in R.” J. Stat. Softw. 15 (9): 1-28.

[48] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. 2018. “Package ’e1071.’” Accessed November 26, 2018. https://cran.r-project.org/web/packages/e1071/e1071.pdf.

[49] Kuhn, M. 2008. “Building Predictive Models in R Using the Caret Package.” J. Stat. Softw. 28 (5): 1-26. doi:10.18637/jss.v028.i05.

[50] R Core Team. 2018. “R: A Language and Environment for Statistical Computing.” R Foundation for Statistical Computing, Vienna, Austria.

[51] Kotsiantis, S. B. 2007. “Supervised Machine Learning: A Review of Classification Techniques.” Informatica. 31: 249-68.

[52] Domingos, P. 2012. “A Few Useful Things to Know about Machine Learning.” Commun. ACM 55 (10): 78-87.

[53] Dietterich, T. 1995. “Overfitting and Undercomputing in Machine Learning.” Comput. Surv. 27: 326-7.

[54] Batista, G., Prati R. C., and Monard, M. C. 2004. “A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data.” ACM SIGKDD Explor. Newsletter 6 (1): 20-9.

[55] Mason, C. H., and Perreault Jr., W. D. 1991. “Collinearity, Power, and Interpretation of Multiple Regression Analysis.” J. Market. Res. 28 (3): 268-80.

[56] Harris, F., Martin, P., and Eagles, H. 2016. “Understanding Wheat Phenology: Flowering Response to Sowing Time.” GRDC Grains Research Update, Wagga, NSW. https://grdc.com.au/Research-and-Development/GRDC-Update-Papers/2016/02/Understanding-wheat-phenology-flowering-response-to-sowing-time.

[57] Myneni, R. B., Keeling, C. D., Tucker, C. T., Asrar, G., and Nemani, R. R. 1997. “Increased Plant Growth in the Northern High Latitudes from 1981 to 1991.” Nature 386: 698-702.
[58] Keyser, A. R., Kimball, J. S., Nemani, R. R., and Running, S. W. 2000. “Simulating the Effects of Climate Change on the Carbon Balance of North American High Latitude Forests.” Global Change Biol. 6 (Suppl. 1): 185-95.

[59] Sturm, M., Racine, C., and Tape, K. 2001. “Increasing Shrub Abundance in the Arctic.” Nature 411: 546-7.