Abstract

Prediction of individual lifespan based upon characteristics evaluated at middle-age represents a challenging objective for aging research. In this study, we used machine learning algorithms to construct models that predict lifespan in a stock of genetically heterogeneous mice. Lifespan-prediction accuracy of 22 algorithms was evaluated using a cross-validation approach, in which models were trained and tested with distinct subsets of data. Using a combination of body weight and T-cell subset measures evaluated before two years of age, we show that the lifespan quartile to which an individual mouse belongs can be predicted with an accuracy of 35.3% (± 0.10%). This result provides a new benchmark for the development of lifespan-predictive models, but improvement can be expected through identification of new predictor variables and development of computational approaches. Future work in this direction can provide tools for aging research and will shed light on associations between phenotypic traits and longevity.

Keywords

aging; classification; longevity; shrunken centroid; T-cell subset; weight

Introduction

Prediction of individual lifespan represents a significant challenge for aging research that is important for understanding factors influencing longevity, as well as identification of lifespan-associated characteristics that can be studied as surrogates of longevity in laboratory experiments. In mice, previous studies have identified early and mid-life traits that correlate with lifespan and have developed models for lifespan prediction (1-11). These investigations have provided insight into links between phenotypic characteristics and mechanisms that regulate longevity. Low body weight, for example, has been associated with increased lifespan in genetically heterogeneous mouse populations (8,11). This association may reflect population-level variability in GH/IGF-I signaling, since mutations inhibiting this pathway have been found to increase lifespan while decreasing body weight in mice (12). Models that predict lifespan also offer the possibility of accelerating aging research. Full-length survivorship studies require several decades to complete in monkeys (13,14), and require at
least three years to complete using healthy mouse strains (15). Forecasting of individual lifespan based upon non-invasive predictors would provide preliminary data years in advance. The prediction of individual longevity, however, represents a daunting task, and is complicated by numerous sources of genetic and environmental variability (16-18). Nevertheless, the potential value of models that accurately predict lifespan is considerable, and efforts directed towards formulating such models will advance our understanding of factors regulating longevity.

There are two complementary areas of research from which progress in lifespan prediction can be anticipated. The first of these involves generation of data sets including a range of predictor variables as well as outcomes of interest, such as lifespan, disease risk, and indices of age-dependent functional decline against candidate predictors can be evaluated. The second element focuses on determining which statistical methods are best able to integrate results from a panel of predictors into a single model. A predictive model represents a pattern connecting a set of quantitative characteristics to individual lifespan. Such patterns, however, may be highly complex and difficult to characterize in simple mathematical terms, e.g. in the form of a regression equation. If presented with a series of facial photographs, for example, we can quickly discriminate young from middle-aged adults with excellent accuracy. This task, however, requires a highly sophisticated pattern recognition device (the human brain), and it is very difficult to write a mathematical model that relates the visual image to the age of the person depicted. Sophisticated modeling approaches may therefore be necessary to characterize patterns that connect predictor variables to longevity and other age-sensitive endpoints of interest. Previous studies have characterized relationships between predictor variables and lifespan using least-squares regression (5-9,11). The advantage of this approach is the strong statistical foundation of least-squares regression procedures, which allows for robust statistical inferences when parametric assumptions are satisfied. A disadvantage, however, is that complex and non-linear relationships may not be adequately captured by regression models. Large datasets, in particular, are valuable for exploring patterns between different traits and lifespan, but are not well-suited to standard regression approaches. The statistical power associated with large datasets may impart statistical significance to small effect sizes of little biological importance. The validity of this statistical inference, moreover, tends to breakdown in large datasets, since multicollinearity among predictor variables accumulates as more variables are incorporated into models (19).

Machine learning algorithms are valuable tools for the generation of predictive models and identification of complex patterns between variables. These methods are used to predict an outcome of interest based upon a set of predictor variables, and are especially well-suited to large datasets for which least-square regression approaches can be problematic. The approach is to present algorithms with a set of training cases, where each training case consists of measurements on a set of predictor variables and a class label. For example, training cases may consist of mice measured for body weight, tail length and body fat percentage (predictor variables), where each mouse is labeled as “short-lived” or “long-lived”. Within this framework, predictive models are generated as the algorithm “learns” a relationship between predictor variables and class labels. The manner by which learning occurs varies among algorithms, but may involve building a decision tree structure, identifying appropriate weights in a neural network, or identifying a maximal margin hyperplane that distinguishes training examples of different classes (20). These approaches have most often been applied in the context of handwritten digit recognition, speech pattern identification and recognition of facial images (21,22). In recent years, however, machine learning has made important contributions to the biological sciences as well, and has been used to annotate gene sequence data (23), diagnose disease (24) and identify compounds with drug activity (25,26). Currently available methods range from classical approaches, such as linear and quadratic discriminant function...
classifiers, to more recently developed methods, which include decision trees, neural networks and support vector machines.

In this paper, we use machine learning to predict mouse lifespan based upon characteristics evaluated prior to two years of age. The dataset we evaluate includes more than one-thousand mice in which a wide range of characteristics have been measured, including T-cell subsets, serum hormone levels (insulin-like growth factor I, leptin and thyroxine), body weight and cataract scores. Previous studies have found these variables to correlate with mouse lifespan using regression-based approaches (5-9,11), and have identified multivariate patterns related to lifespan (6). In this report, we evaluate the predictive capacity of variables using machine learning and a stringent evaluation criterion based upon cross-validation. In this approach, machine learning models are constructed and evaluated using non-overlapping sets of observations (drawn at random from the same mouse population). We can in this way evaluate how accurately mouse lifespan is predicted when models are applied to new animals not used in model construction. This provides an assessment of model “generalization ability”, which is of great importance for determining the usefulness of models in practical contexts (20). The framework we present provides a new perspective on the use of phenotypic characters as prognostic tools in aging research, and our results represent a novel benchmark for future work aimed at prediction of lifespan.

Methods

Experimental Animals

Data were collected from a genetically heterogeneous stock of mice generated by crossing (BALB/cJ × C57BL6/J)F1 (CB6F1) females with (C3H/HeJ × DBA/2J)F1 (C3D2F1) males. Mice derived from this cross are genetically unique, but share 50% of genes with other members of the population. The genetic composition of this stock thus minimizes the chance that findings will apply only to a single inbred or F1 hybrid genotype (15,27). Standard maintenance protocols used for this stock have been described previously (7,11). In brief, mice were weaned at 19 - 21 days of age and maintained in cages containing 3-4 mice of the same sex. Non-virgin (“mated”) female mice were produced by introducing a male into cages containing 3 - 4 female mice when the females were 2 months old. Litters were then removed from these cages within the first week after birth, and the male removed when the females were 6 months old. Throughout the study, free access to laboratory chow and fresh water supply was provided, and sentinel animals were exposed to spent bedding on a quarterly basis to check for possible pathogen infection. Three animals were euthanized following detection of mouse parvo virus, but no evidence of pinworm, Sendai virus, mycoplasma or coronavirus was obtained during the study period. All cages were inspected on a daily basis to evaluate health of the animals and to check for mortalities. Mice found to be extremely ill were euthanized, and the date of euthanasia was recorded as the most likely date of natural death.

Lifespan Predictor Variables

Table 1 describes measurements obtained during the study that serve as potential lifespan-predictors in the present analysis. Variables include T-cell subset scores evaluated at 8 and 18 months of age, serum hormone levels at 4 and 15 months of age, body weight at 8 - 18 months of age and cataract scores at 18 and 24 months of age. Several predictor variables listed in Table 1 have previously been identified as significantly correlated with mouse lifespan. Miller (6) showed that, within individual gender categories, variables CD4_8, CD8M_8, CD4V_8, CD4_18, CD4M_18, CD8M_18, CD4V_18 and CD4P_18 were significantly correlated with lifespan. A later investigation revealed that combining these variables using principal component analysis provided an index with an even stronger association with longevity (7). Miller et al. (8) focused on body weight measurements, and found that body weight between
the ages of 2 and 24 months was correlated with lifespan, such that smaller mice were long-lived as compared to larger mice. Harper et al. (9) investigated hormone levels as potential predictors of lifespan, and reported that LEP_4, T4_4 and IGF_15 were correlated with both body weight and lifespan. Surprisingly, Harper et al. (9) also found significantly greater longevity among mice with more advanced cataract formation (higher Cat24 scores). Harper et al. (11) confirmed, in a replicate cohort, that CD8M_8, CD8M_8 and CD4_18 were significantly correlated with longevity, and also found that CD8_18 and CD8P_18 were associated with lifespan. In addition, Harper et al. (11) reported that combination of T-cell subset and body weight measures significantly increased the variation in lifespan that was explained by regression models, compared to models based only on T-cell or weight predictors individually.

Data Preprocessing

The complete dataset includes age at death and predictor variable measurements for 1188 mice (403 virgin males, 457 virgin females, 299 mated females). The data are pooled from two survivorship studies that have been referred to as LAG1 and LAG2 in previous publications (11). We eliminated 183 cases in which mice had died prematurely from mouse urinary syndrome, or had been removed from colonies due to fighting. Additionally, mice dying within the first two years of life were not considered in the analysis. This was done because all predictor variables were recorded within the first two years of life and it was of interest to predict the longevity of mice surviving beyond this two-year time frame. After this preprocessing, the dataset contained 741 mice, all of which died of apparently natural causes beyond two years of age.

The 741 mice were each assigned to a lifespan quartile based upon age at death, and lifespan quartile was used as the primary response variable in all analyses. This was done because many machine learning algorithms have been designed to predict categorical, rather than continuous outcomes. Focusing on lifespan quartiles therefore allowed us to implement a wider range of machine learning techniques. In addition, predicting categorical outcomes allows success to be measured in terms of an accuracy or error rate, which is useful for judging the performance of algorithms, as well as the quality of predictor variable subsets (20). For virgin males, lifespan quartiles were assigned based upon a 25th percentile lifespan of 811 days, a 50th percentile of 890 days, and a 75th percentile of 1034 days. Among virgin females, corresponding 25-50-75 percentile lifespans were 811, 892 and 978 days, respectively. For mated females, these values were 790, 855 and 954 days, respectively.

All continuous variables were standardized to have a mean of zero and variance of one, and in some cases were log-transformed to attenuate the influence of outliers. Since a complete set of predictor variable measurements was not available for all mice, it was necessary to impute for missing data. For the discontinuous variable “LitSize”, we calculated the most commonly occurring value among all mice, and substituted this value in place of missing data. All other variables with missing data were continuous. In these cases, we calculated the median value among all observations, added a small amount of random noise, and substituted the obtained value for missing data. Substitution with median values is one of several possible approaches for dealing with missing data (28). The rationale is that variable values further away from the median are more influential in generating predictive models, such that substitution of the median value is neither helpful nor unhelpful for lifespan prediction.

Machine Learning and Lifespan Prediction

Machine learning methods vary considerably in their complexity and it is uncertain for any particular application which algorithm will provide the best overall performance (20). The best algorithm depends on the quality of predictor variables and the “decision boundaries” that best
distinguish mice belonging to different lifespan quartiles (Figure 1). If quartiles are clearly separated by linear decision boundaries, simple algorithms may predict lifespan quartile with high accuracy (see Figure 1A). On the other hand, if quartiles can only be separated by complex, irregular decision boundaries, accurate prediction of lifespan quartile will require more sophisticated algorithms (see Figure 1B). We therefore applied many algorithms in order to determine which approaches provide best results for prediction of lifespan quartile (Table 2). The manner by which algorithms generated predictions varied considerably among the different methods. Some algorithms generated predictions by constructing regression models (simple linear regression, multinomial logistic regression), while others constructed decision trees or sets of classification rules (e.g., binary decision trees, random forest, Ripper rule learner). Other algorithms fit probability models to mice associated with each lifespan quartile, and used posterior class probability as a basis for prediction (e.g., Naïve Bayes). In some cases, predictions were generated in non-parametric fashion without construction of a formal classification model (e.g., k-nearest neighbor, nearest centroid, nearest shrunken centroid). All algorithms were implemented using the R statistical software package. The R extension package and specific function used to implement each algorithm is listed in Table 2, and full descriptions of each package and function can be accessed online (http://www.r-project.org/).

If predictive models are trained and evaluated using the same data, it is always possible to generate a model that predicts with 100% accuracy. This can be done, for example, by constructing an overfit regression model with a large number of parameters. The predictive performance of such a model, however, would be very sensitive to chance variation or noise in the training samples chosen to construct the model. Consequently, the model would lack “generalization ability” and not be useful as a prognostic tool (20). The ability of an algorithm to produce generalizable results can be tested by using cross-validation to evaluate prediction accuracy, a process in which the model is trained using one set of data, and its prediction accuracy evaluated using a another, distinct set of data. This evaluation procedure is commonly used in the development of predictive models, and is necessary for providing a realistic assessment of how useful a model can be in practical scenarios, where the interest is to forecast results based upon predictor variables from new observations (e.g., partially complete survivorship studies). Our analysis therefore used cross-validation to evaluate the accuracy of each algorithm for prediction of lifespan quartile. Approximately 90% of the 741 mice were randomly selected as “training data”, with the remaining 10% of mice serving as “testing data”. Machine learning models were constructed using only the training data, and the resulting model was then applied to the testing data to determine prediction accuracy. The process of randomly parsing observations into training and testing sets, training the algorithm, and applying the algorithm to test data was repeated 10,000 times for each algorithm evaluated. Training datasets were always constructed such that an equal number of cases from each lifespan quartile were presented to the algorithm. Testing datasets contained approximately 20 cases from each lifespan quartile. The average prediction accuracy attained by each algorithm was compared to that expected from random guessing (25%). A 95% confidence interval was calculated for each average prediction accuracy and used to evaluate whether accuracy was significantly above 25%, or whether performance differed significantly among alternative algorithms.

Results

Our analysis considers only mice that lived beyond two years of age, but relationships between predictor variables and lifespan were found that resembled those identified in previous analyses (5-9,11). For each variable, we used two-sample t-tests to evaluate whether significant differences existed between mice assigned to the shortest and longest-lived lifespan quartiles. This analysis revealed robust relationships between lifespan and body weight measures (W8, W10, W12 and W18), which were generally lower in mice belonging to the longest-lived quartile. This association was significant when genders were pooled (P < 0.015), or among
males alone ($P < 0.001$), and the same trend was present (but non-significant) amongst females and mated females ($0.073 < P < 0.737$). At 15 months of age, mice belonging to the longest-lived quartile had reduced leptin levels ($P = 0.05$; genders pooled), as well as lower serum IGF-I ($P = 0.045$; males only). With regard to T-cell subsets, mice belonging to the longest-lived quartile had significantly lower scores on the variables CD4M_8, CD8M_8 and CD8M_18 ($P < 0.04$; genders pooled), and significantly higher scores on CD4_18 and CD4V_18 ($P < 0.02$; genders pooled). There was, however, considerable scatter in these associations, and no pair of predictors consistently distinguished between mice from the shortest and longest-lived quartiles (Figure 2A and 2B). Linear discriminant function and principal component analyses were used to reduce the 29 continuous predictors to two dimensions. For both approaches, however, only slight separation between lifespan quartiles was attained, and there was no simple decision boundary that distinguished between classes (Figures 2C and 2D).

**Comparison of Machine Learning Algorithms**

Predictive models were constructed using each of 22 machine learning algorithms (Table 2). For each algorithm, models were generated using all 31 predictor variables, as well as subsets of 12 and 6 predictor variables that were chosen by ranking variables according to their importance to prediction accuracy based on the Random Forest algorithm (29) (see Figure 3). This algorithm generates a useful measure of variable importance based upon cross-validation and the accuracy decline that results when predictors are removed from the dataset (30). We used this measure to rank all 31 predictor variables in terms of their importance for predicting lifespan quartile (Figure 3). This suggested that the subset CD4V_18 and body weight measurements were the most important variables for lifespan quartile prediction, while some variables, such as T4_4, T4_15 and LEP_15, may have a net deleterious effect on prediction accuracy. These results were supported by our evaluation of algorithm performance (Table 2), since most algorithms performed best when only the 12 or 6 most important predictors were used (Table 2). For instance, using all 31 predictor variables, support vector machine (SVM) predicted lifespan quartile with 29.90% accuracy, but this value increased to 34.03% when only the best 12 predictor variables were used.

All algorithms predicted lifespan quartile with accuracy significantly greater than that expected by random guessing (25%), and in some cases, simple approaches performed as well as highly complex methods (Table 2). Regression-based approaches, which have been used previously to generate lifespan-predictive models (e.g., 8, 9, 11), performed well in some cases (31.8% accuracy), but were not among the best-performing algorithms (see Table 2; multinomial logistic regression and simple linear regression). Overall, three algorithms emerged as the most promising (Nearest Shrunken Centroid (NSC), SVM and Stabilised Linear Discriminant Analysis (SLDA)). SVM and SLDA obtained a maximal accuracy of 34.03% and 34.06%, respectively, using the 12 most important predictor variables (Table 2). The NSC algorithm, however, attained slightly greater accuracy using only six predictors (34.37%).

**The NSC Algorithm: Variable Selection**

Further experiments were performed to determine whether NSC could attain greater accuracy using more than six predictors, or attain a similar level of accuracy with fewer predictors. An initial model was constructed using only the two most important predictors (CD4V_18 and W10), which yielded an average accuracy of 33.1%. Variables were then added individually according to their importance (Figure 3). We found that accuracy peaked at 34.5% following the addition of five predictor variables (CD4V_18, W8, W10, W12 and W18) (Figure 4). This analysis suggested that strong performance could be attained using a relatively small number of predictor variables.
We evaluated the prediction accuracy attained by the NSC algorithm using each of the $\binom{31}{2} = 465$ possible predictor variable pairs (at least 100 cross-validation simulations were performed per pair). This analysis indicated that CD4V\_18 and W18 provided the best two-variable subset, which yielded a prediction accuracy of 33.5%. The second best two-variable subset provided 33.4% accuracy and also included CD4V\_18, but in combination with another T-cell measure (CD8M\_8) rather than a body weight variable. We next evaluated each of the $\binom{31}{3} = 4495$ three-variable subsets (using at least 20 simulations per subset), and found that best performance was attained with CD4V\_18, W8 and W18 as predictors (34.4%). Taken together, these results suggested that the best four-variable subset would include CD4V\_18. We thus evaluated the $\binom{30}{3} = 4060$ possible four-variable subsets that include CD4V\_18, and found that a subset of CD4V\_18, W8, W10 and W18 yielded an accuracy of 34.6%. Continuing in this fashion, we examined the $\binom{29}{3} = 3654$ possible five-variable subsets that include CD4V\_18 and W18. This analysis suggested a five-variable model with three body weight variables (W8, W10 and W18) and two T-cell subset variables (CD4V\_18 and CD4M\_8), which yielded an accuracy of 35.3% (95% CI: 35.2% - 35.4%). This was the best overall performance in our analysis, as further exhaustive searches did not identify a six-variable model with significantly improved accuracy. We also found that inclusion of principal components that combined data from multiple weight measures or related sets of T-cell subsets did not improve the accuracy of the model. The performance of our best overall predictive model is illustrated by Figure 5, which displays the distribution of accuracies obtained among 10,000 cross-validation simulations, and provides a comparison to the distribution of accuracies obtained by an algorithm that generates predictions randomly.

The NSC Algorithm: Additional Analyses

The five-variable prediction model we developed was used to separate mice into groups that, based upon five variables measured before two year of age, are projected to be either shorter or longer-lived. For each of the 741 mice considered in our analysis, a predicted lifespan quartile was generated based upon all other 740 mice (leave-one-out cross-validation). We then evaluated mice assigned to the two shorter-lived lifespan quartiles as compared to the two longer-lived quartiles. Mice projected to be longer-lived had a mean lifespan of 930.6 days, while mice projected to be shorter-lived had a mean lifespan of 879.5 days. Survivorship curves associated with these two groups were significantly different from each other ($P < 0.0001$; log-rank test) (see Figure 6), and mice projected to be longer-lived had age-specific mortality rates reduced significantly by 36.9% (95% CI: 26.8% - 45.7%) ($P < 0.0001$; proportional hazard model).

Mice projected to be longer-lived also had significantly greater maximum lifespan, using survival to the pooled 90th percentile as a criterion for extreme longevity. Among all mice, the 90th percentile lifespan was 1087 days, and while only 3.6% of mice projected to be shorter-lived survived this long, we found that 15.0% of mice assigned to the longer-lived group survived this time period ($\chi^2 = 26.45; P < 0.0001$).

The estimated probability of a mouse belonging to quartile $x$ given its scores on a number of variables is referred to as posterior class probability (20). Posterior class probabilities are generated by many machine learning methods and provide a useful form of continuous output. Such output represents an integration of potentially many predictor variables, and in the present context, can be used to construct an index associated with lifespan. Using the leave-one-out method described above, we calculated the posterior probability of belonging to the longest-lived quartile for each of the 741 mice considered in our analysis, based upon the NSC algorithm and the probabilistic framework developed by Tibshirani et al. (31). This provided an index that was significantly associated with lifespan ($r = 0.15, P < 0.001$) (Figure 7). Likewise, we calculated the posterior probability of belonging to the shortest-lived quartile for each mouse,
which provided an index exhibiting the opposite pattern of association with lifespan ($r = -0.14; P < 0.001$). These associations were not dependent upon outlying observations, since the same trends were identified based upon the Spearman rank correlation ($|r_s| > 0.14, P < 0.001$).

Prediction accuracy increases systematically as more training examples are used during the algorithm learning phase. At some point, however, additional training examples provide only negligible returns in predictive accuracy. In our analysis, training was performed with randomly sampled sets of 664 mice (90% of all mice that met our inclusion criteria). To determine whether this training set size exceeds a point of diminishing returns, we varied the number of training examples between 76 and 664 mice, and used cross-validation to evaluate predictive accuracy for each number of training examples (using NSC and the five variable model developed above). Accuracy gains due to increased training set size diminished slightly beyond 400 training samples (Figure 8). A moving average time series model was used to forecast the accuracy expected for 1000 training examples, based upon the observed increase in accuracy between 76 and 664 training examples. This analysis indicated that, based on the set of five predictors from which our best model was developed, 1000 training samples would yield an average accuracy of 40.0% ($± 1.2\%$) (Figure 8).

Discussion

Models that predict individual lifespan can provide a valuable tool for aging research, and development of such models represents a platform for understanding relationships between longevity and other phenotypic characteristics. Full-length survivorship experiments require years to complete and are a rate-limiting step in the study of mammalian aging. Clearly, such experiments will never be replaced by predictive models. Well-developed models, however, can generate preliminary data years in advance, and the output of such models, as an integration of multiple variables that may predict lifespan individually, can provide a surrogate target for aging research that is easier to evaluate than lifespan. Indeed, a similar concept has already been successfully applied using the Drosophila model system (32). Using gene expression patterns that predict lifespan, Baur et al. (32) developed a survivorship screen that satisfactorily revealed the known effects of temperature, caloric restriction and resveratrol on Drosophila lifespan, but required 80% less time than a full-length survivorship study. We have here explored the plausibility of such a model in mice, and our results provide a realistic indication of how accurately mouse lifespan can be predicted using state-of-the-art machine learning algorithms. Using a set of physiological measurements made prior to 24 months of age, we have shown that mouse lifespan quartile can be predicted with an average accuracy of 35.3%. This estimate is based upon a stringent cross-validation criterion, in which accuracy is based upon prediction of test cases not used in model construction. We suggest that considerable potential remains for improving upon the 35.3% prediction accuracy attained in this study.

Development of new computational approaches, generation of larger datasets, and identification of candidate predictor variables are all avenues by which our result can be improved.

Machine learning approaches vary considerably in terms of their complexity, performance on easy versus difficult tasks, and the ease with which investigators can apply these methods. We have carried out a performance evaluation of many machine learning approaches for the specific task of predicting mouse lifespan, which provides guidance for future studies addressing this issue. It should be borne in mind that, for any algorithm, performance is data-dependent and will vary across contexts. It is therefore impossible and incorrect to identify a single “best algorithm” that will always yield the best predictive performance (20). We anticipate, however, that any survivorship dataset containing lifespan-predictor variables will share some properties with data analyzed here. For instance, some level of “overlap” between mice belonging to different lifespan quartiles seems inevitable, such that accurate predictions
will require recognition of complex decision boundaries (see Figures 1 and 2). The best algorithms identified in our analysis were most successful at characterizing such decision boundaries, and therefore provide good starting points for future investigations aimed at prediction of mouse lifespan. We found that lifespan was accurately predicted by support vector machine and stabilised LDA algorithms, but that the Nearest Shrunken Centroid approach provided the best overall performance. This algorithm was developed in the context of DNA microarray datasets and used to classify cancer subtypes based upon gene expression patterns (31). The approach was also straightforward to implement and can be understood by non-experts. It was computationally inexpensive, as 10,000 simulations could be carried out in less than one hour on a standard desktop PC; other algorithms listed in Table 2 required as much as two days for execution using the same hardware. The method is similar to the Nearest Centroid approach, which uses \( p \) predictor variables to obtain a \( p \)-dimensional centroid for each lifespan quartile, and predicts lifespan quartile for new cases by determining which centroid the test case is most similar to. In the Nearest Shrunken Centroid approach, however, class centroids are standardized and “shrunk” towards an overall centroid (calculated by averaging centroids among lifespan quartiles), which has the effect of emphasizing those predictors with low levels of variation in each class (see 31). This yields a classification rule characterized by complex, non-linear decision boundaries, which has been found to perform well on challenging machine learning tasks (31,33).

The performance of machine learning approaches can generally be improved by providing a larger number of training samples during the algorithm learning stage. Numerous sources of variation obscure relationships that exist between lifespan quartile and the predictor variables we measured. To deal with this variation, effective algorithms must accurately characterize underlying biological relationships, and do this consistently despite “noise” that obscures underlying relationships. In other words, for accurate prediction, decision boundaries chosen by algorithms should be unbiased, but at the same time, must be robust to noise and chance variation within training data. In general, these aspects of predictive modeling are in tension, but by increasing the number of training samples for algorithm learning, it is often possible to decrease both the bias and variance associated with decision boundaries (20). Our study evaluated models based upon 664 training samples, which is small compared to training set sizes commonly used for machine learning tasks. In the context of handwritten digit recognition, for instance, benchmark tasks commonly include more that 50,000 training samples (21,34). We therefore expect that the generation of larger datasets would lead to further improvements in lifespan prediction accuracy. We achieved 35.3% prediction accuracy using 664 training samples, but our analyses indicate that 1000 training samples would yield an accuracy of approximately 40% (Figure 8) with the same, limited set of predictors. This suggests that a survivorship experiment generating 1000 mice (living beyond two years of age) would provide a more useful model than that developed in this study. Even larger experiments could only improve predictive accuracy, but risk low returns in exchange for substantial efforts and costs.

The successful prediction of mouse lifespan ultimately depends upon the quality of predictor variables that are available. Predictor variables, or some multivariate combination of predictors, must have a linear or non-linear association with lifespan. Otherwise, a useful model cannot be constructed, regardless of the computational approach that is taken. The basis of associations between predictors and lifespan is of special interest. Predictor variables can be associated with lifespan through a specific disease process. For example, if leukemia is a major factor contributing to mortality in a cohort, blood cell counts may provide an effective predictor of longevity, and the overall predictive model will be uninformative with regard to aging mechanisms. Alternatively, predictor variables can be associated with lifespan because they characterize some aspect of the aging process. We found, for example, that body weight measurements between 8 and 18 months of age were among the most important predictors of
lifespan quartile, and that four of the top five predictors were body weight variables. These results are consistent with our findings from earlier studies (8,11), as well as multiple investigations showing that, within individual species, lower body weight is commonly associated with increased longevity (35-37). Links between body weight, aging and longevity are not fully understood, but the GH/IGF-I axis may be an underlying explanatory factor (12), and in fact, we found that among mice living beyond two years of age, body weight values were correlated with serum IGF-I levels at 4 and 15 months of age ($0.42 < r < 0.46$). In some segregating mouse populations, low IGF-I levels early in life do indeed predict longer lifespan (38).

Identification of new phenotypic traits that predict lifespan may be the most promising approach for improving upon our results. We examined 31 predictor variables, but lifespan was most accurately forecasted when only the five strongest predictors were used in model construction. The strongest predictors in our current dataset are drawn from two physiological domains, body weight and T cell biology, and there are strong correlations among the various body weight measures as well as among the T cell subset frequencies. Including several measures of weight, or several measures of T cell status, may improve the performance of these predictive algorithms by diminishing the variation reflecting measurement error or day-to-day variations in biological state that are independent of inter-individual differences (for example, the effects of a recent meal or hydration state on measured body weight). We suspect that incorporation of measures drawn from other physiological domains, such as tests of motor function, cognitive function, or resistance to DNA damage, might be a useful step towards improved predictive performance. Previous studies, for example, have identified other non-invasive phenotypic traits that are correlated with lifespan, which were not evaluated in our study. Harrison and Archer (2), for example, showed that tight wire clinging ability, open field activity, collagen denaturation rate, hair regrowth, wound healing and blood hemoglobin concentration were all correlated with longevity in certain mouse genotypes. Flurkey et al. (15) described several other potentially useful phenotypic traits, such as urine concentrating ability and carbon dioxide production, for which relationships with longevity have not been examined. A nearly endless list of possible predictor variables is provided by expression measures of individual genes, or composite measures derived from a number of genes. In skin tissue, for instance, $p16^{INK4a}$ has been found to correlate with chronological age, although its value as a predictor of longevity has not been previously considered (39,40). Some gene expression variables may reflect certain physiological states beneficial for longevity, and may therefore prove to be unexpectedly valuable longevity predictors, even when population-level variation is attributable to non-genetic or “chance” sources in uniform environments (e.g., 41). For genetically heterogeneous populations, such as that considered in the present study, genetic polymorphisms represent an especially important avenue for investigation. Various forms of genetic data, such as single nucleotide polymorphisms, could readily be incorporated into models as categorical predictor variables, which may improve prediction accuracy by accounting for interactions between genotypic and phenotypic characters. Such interactions may be of considerable importance, since Harrison and Archer (2) found that relationships between phenotypic traits and longevity were often dependent upon genotype.

The methods employed in the present study also provide a useful framework for investigations aimed at discovering and validating biomarkers of aging. Aging biomarkers provide an operational definition of aging that is useful for testing hypotheses about aging experimentally, such as whether aging is delayed by mutations or environmental interventions (15). The denaturation rate of collagen, for example, has been used as a biomarker to suggest that aging is delayed in long-lived mouse strains (42). The quest for aging biomarkers has been controversial, and hindered by both conceptual and methodological hurdles, including disagreement regarding the definition of “biomarker”, overlap between aging and disease processes, and lack of knowledge about aging mechanisms (43-46). For example, there is an
important conceptual distinction between predictors of lifespan and biomarkers of aging. To be useful as a biomarker of aging, a measurement must be able to distinguish among individuals who are aging at different rates, for example because of differential exposure to a putative antiaging intervention. Conceptually, such traits must be measured in individuals who have already experienced some effects of the aging process, for example among a group of middle-aged adults. Inherited alleles, or weight at birth, for example, may prove to be strong predictors of lifespan or of other age-related endpoints, and yet have no value as measures of the rate at which aging has occurred in a specific individual.

Nevertheless, predictive capacity is an important component of biomarker validation, and it is widely agreed that legitimate biomarkers must predict the outcome of a wide range of age-sensitive events in different physiological systems, and ultimately, should be predictive of individual lifespan (2,4,47-50). The cross-validation strategy used in the present study, involving random division of data into training and testing sets, provides a useful way of evaluating predictive capacity and thus provides a framework for biomarker validation. Furthermore, we have used machine learning to generate posterior class probabilities for different lifespan quartiles (see Figure 7). Such posterior probabilities may provide useful indices that themselves serve as biomarkers or as indicators of “biological age” (51-53). A key advantage of this approach is that alternative models used to generate posterior probabilities can be compared based upon cross-validation lifespan-prediction accuracy. This could make progress stemming from the biomarker research agenda more transparent from skeptical viewpoints, since for example, it would be unambiguous to determine whether an index generated by predictors $X$, $Y$ and $Z$ is superior to an index generated by predictors $P$, $Q$ and $S$.

The model we developed predicts lifespan quartile with 35.3% accuracy, and we have shown that this model can assign mice to groups that differ in mean lifespan by 5.8% (930.6 versus 879.5 days; see Figure 6). At a glance, this 5.8% difference seems minor. It should be noted, however, that an algorithm predicting death from cancer in the American population (with 100% accuracy) would separate humans into groups differing by less than 3% in mean lifespan (54). We therefore suggest that our model provides informative predictions with biologically meaningful underpinnings. Furthermore, we are optimistic that more accurate lifespan-predictive models can be formulated in future investigations. The computational approach that performed best in our analysis was developed within the last decade (31). In the coming decade, intensive research efforts will focus on improving variable selection methods (55), extracting informative features from a set of variables (56), ensemble approaches for combining outputs of multiple algorithms (57), and filtering of training samples to identify maximally informative cases for learning (58). Such investigations will lay the groundwork for development of new algorithms, which can be complemented by better understanding of aging mechanisms and exploitation of lifespan-predictors at the molecular and phenotypic levels. These approaches, both computational and biological, can improve the accuracy with which mouse lifespan is predicted and should advance interdisciplinary approaches to biogerontology and the study of mammalian aging.

Acknowledgements

We thank Maggie Lauderdale and Jessica Sewald for technical and husbandry assistance. We also thank two anonymous reviewers for helpful comments on this manuscript. This work was supported by grants AG11687 (RAM), AG024824 (RAM) and the University of Michigan Department of Pathology. WRS is supported by NIA training grant T32-AG00114.

References


Figure 1.
Simple and complex relationships between predictor variables and lifespan. Each plot displays data for twelve subjects in relation to scores on two predictors. The four symbol types represent subjects associated with each lifespan quartile. (A) Simple relationship between predictors and lifespan. Subjects associated with different lifespan quartiles are distinguished by simple linear decision boundaries indicated by dotted lines. Simple learning algorithms may perform well in comparison to complex algorithms. (B) Complex relationship between predictors and lifespan. Subjects associated with different lifespan quartiles can only be distinguished through recognition of irregular decision regions. Complex learning algorithms are required for accurate prediction of lifespan based on the two predictors.
Figure 2.
Overlap between short and long-lived lifespan quartiles. In each plot, open circles (○) represent mice associated with the shortest-lived quartile, while plus signs (+) represent mice associated with the longest-lived quartile. Plots (A) and (B) display quartile relationships with respect to body weight and the CD8M T-cell subset at 8 and 18 months of age, respectively. In (C), all mice in the dataset have been plotted with respect to two linear discriminant axes, which were derived from the 29 continuous predictor variables listed in Table 1. In (D), mice are plotted with respect to the first two principal components derived from the 29 continuous predictor variables. The two principal components account for 6.8% of the total variation among the 29 variables.
Random Forest evaluation of variable importance. Variables listed near the top of the figure are most important for prediction of lifespan quartile. The Random Forest algorithm constructs a “forest” of decision trees, where each tree generates predictions based upon a randomly chosen set of predictor variables. Predictions are then generated by majority vote among all trees in the forest (29). For the 741 mice we considered, the algorithm generates forests using a bootstrap sample of approximately 500 mice and remaining mice are used as an internal testing set for accuracy evaluation. For a given predictor $X$, predictive accuracy is evaluated before and after permuting values of $X$ among mice. The difference between the two obtained accuracies is then determined and used as a measure of variable importance. This procedure was repeated 10,000 times using forests of 2000 decision trees per trial. The figure shows the average accuracy change among all 10,000 trials.
Figure 4. Variable subset evaluation. Prediction accuracy was evaluated for different variable subsets using the Nearest Shrunken Centroid algorithm. Variables were entered into the model according to their importance, as indicated by results shown in Figure 3. The first model included variables CD4V_18 and W10 as lifespan predictors. The accuracy obtained using this model was evaluated by ten-fold cross-validation and 10,000 simulations. The average accuracy among all simulations is represented by the left-most bar shown in the plot. Each subsequent bar from left to right indicates the accuracy obtained by adding the next most important variable to the model. The right-most bar represents the mean accuracy obtained using the full model with all 31 predictor variables.
Figure 5.
Simulation results. The Nearest Shrunken Centroid (NSC) algorithm was implemented using five predictor variables (CD4M_8, CD4V_18, W8, W10 and W18). The plot shows the distribution of accuracies obtained on the testing set among 10,000 simulation trials. The solid density represents the accuracy distribution obtained by a random guessing algorithm, while the dotted line represents the accuracy distribution obtained by the NSC algorithm. The solid vertical line indicates the mean accuracy among all trials for the random guessing algorithm (approximately 25%). The dotted vertical line indicates the mean accuracy among all trials for the NSC algorithm (35.3%).
Figure 6.
Predicted lifespan and survivorship. The lifespan of each mouse in the data set was predicted using Nearest Shrunken Centroid and the leave-one-out method (see text). Five predictors were used in the model (CD4M_8, CD4V_18, W8, W10 and W18). The dotted line represents the cohort of mice with predicted lifespan in the upper two quartiles, while the solid line represents the cohort of mice with predicted lifespan in the lower two quartiles. The dotted vertical line indicates the minimum lifespan of mice that were included in the analysis (720 days). Zero survivorship is indicated by the dotted horizontal line.
Figure 7.
Posterior probability and lifespan. The posterior probability of belonging to the longest-lived quartile was determined for each mouse in the dataset using the leave-one-out method (see text). Five predictors were used in the model (CD4M_8, CD4V_18, W8, W10 and W18). The dotted line represents a least squares regression line.
Figure 8.
Effect of training set size on accuracy. The number of training samples was varied between 10% and 90% of the total dataset (between 75 and 667 mice). For each training set size, the mean accuracy obtained by the Nearest Shrunken Centroid algorithm was determined by 10000 simulations (using ten-fold cross-validation). The mean accuracies obtained for each training set size are indicated by the solid line. Dashed lines represent forecasted accuracy for larger training set sizes (between 667 and 1000 mice). The middle dashed line indicates the forecasted accuracy, while the upper and lower dashed lines represent standard error margins. Forecasts were generated using a moving average model with two parameters and two degrees of differencing. Model selection was based upon the AIC criterion described by Brockwell and Davis (72).
Table 1
Predictor variables. The table lists the age at which data were obtained and provides a brief description for each variable. See references (5-9) and (11) for further description of variables.

<table>
<thead>
<tr>
<th>Variable ID</th>
<th>Age (months)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-Cell Subsets</td>
<td>CD3_8</td>
<td>8</td>
</tr>
<tr>
<td>T-Cell Subsets</td>
<td>CD3_18</td>
<td>18</td>
</tr>
<tr>
<td>T-Cell Subsets</td>
<td>CD4_8</td>
<td>8</td>
</tr>
<tr>
<td>T-Cell Subsets</td>
<td>CD4_18</td>
<td>18</td>
</tr>
<tr>
<td>T-Cell Subsets</td>
<td>CD4M_8</td>
<td>8</td>
</tr>
<tr>
<td>T-Cell Subsets</td>
<td>CD4M_18</td>
<td>18</td>
</tr>
<tr>
<td>T-Cell Subsets</td>
<td>CD4P_8</td>
<td>8</td>
</tr>
<tr>
<td>T-Cell Subsets</td>
<td>CD4P_18</td>
<td>18</td>
</tr>
<tr>
<td>T-Cell Subsets</td>
<td>CD4V_8</td>
<td>8</td>
</tr>
<tr>
<td>T-Cell Subsets</td>
<td>CD4V_18</td>
<td>18</td>
</tr>
<tr>
<td>T-Cell Subsets</td>
<td>CD8_8</td>
<td>8</td>
</tr>
<tr>
<td>T-Cell Subsets</td>
<td>CD8_18</td>
<td>18</td>
</tr>
<tr>
<td>T-Cell Subsets</td>
<td>CD8M_8</td>
<td>8</td>
</tr>
<tr>
<td>T-Cell Subsets</td>
<td>CD8M_18</td>
<td>18</td>
</tr>
<tr>
<td>T-Cell Subsets</td>
<td>CD8P_8</td>
<td>8</td>
</tr>
<tr>
<td>T-Cell Subsets</td>
<td>CD8P_18</td>
<td>18</td>
</tr>
<tr>
<td>Hormones</td>
<td>T4_4</td>
<td>4</td>
</tr>
<tr>
<td>Hormones</td>
<td>T4_15</td>
<td>15</td>
</tr>
<tr>
<td>Hormones</td>
<td>LEP_4</td>
<td>4</td>
</tr>
<tr>
<td>Hormones</td>
<td>LEP_15</td>
<td>15</td>
</tr>
<tr>
<td>Hormones</td>
<td>IGF_4</td>
<td>4</td>
</tr>
<tr>
<td>Hormones</td>
<td>IGF_15</td>
<td>15</td>
</tr>
<tr>
<td>Body Weight</td>
<td>W8</td>
<td>8</td>
</tr>
<tr>
<td>Body Weight</td>
<td>W10</td>
<td>10</td>
</tr>
<tr>
<td>Body Weight</td>
<td>W12</td>
<td>12</td>
</tr>
<tr>
<td>Body Weight</td>
<td>W18</td>
<td>18</td>
</tr>
<tr>
<td>Other</td>
<td>LitSize</td>
<td>N/A</td>
</tr>
<tr>
<td>Other</td>
<td>Cat18</td>
<td>18</td>
</tr>
<tr>
<td>Other</td>
<td>Cat24</td>
<td>24</td>
</tr>
<tr>
<td>Other</td>
<td>Gender1</td>
<td>N/A</td>
</tr>
<tr>
<td>Other</td>
<td>Gender2</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Table 2
Algorithm Performance Summary. Lifespan quartile prediction accuracy was evaluated for 22 machine learning algorithms using all 31 predictor variables, the 12 most important predictors, and the 6 most important predictors. Variable importance was determined based upon the Random Forest algorithm (see Figure 4). Algorithms are ranked according their best overall performance among the three predictor variable subsets. For each listed value, accuracy is based on 10,000 simulations in which 664 mice (90%) were randomly selected as training data and used in model construction, with 77 mice (10%) used as testing data for model evaluation (see Methods). For each simulation, the average number of correct lifespan quartile predictions among 77 testing set mice was determined. The table lists the average percent accuracy obtained among all 10,000 simulations (95% confidence intervals are approximately ± 0.10%). References and/or descriptions of each algorithm are provided below the table, along with parameter settings where applicable. The R package and function used to implement each algorithm is given in brackets [package, function].

<table>
<thead>
<tr>
<th>Method</th>
<th>31 Predictors</th>
<th>12 Predictors</th>
<th>6 Predictors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nearest Shrunken Centroid&quot;</td>
<td>32.86</td>
<td>33.61</td>
<td>34.37</td>
</tr>
<tr>
<td>Stabilised Linear Discriminant Analysisb</td>
<td>30.50</td>
<td>34.06</td>
<td>32.30</td>
</tr>
<tr>
<td>Support Vector Machinec</td>
<td>29.90</td>
<td>34.03</td>
<td>33.57</td>
</tr>
<tr>
<td>Gaussian Processd</td>
<td>30.29</td>
<td>32.81</td>
<td>33.90</td>
</tr>
<tr>
<td>Conditional Inference Tree Foreste</td>
<td>29.88</td>
<td>33.49</td>
<td>32.65</td>
</tr>
<tr>
<td>Random Forestf</td>
<td>30.18</td>
<td>33.34</td>
<td>29.36</td>
</tr>
<tr>
<td>Support Vector Machineg</td>
<td>29.56</td>
<td>33.27</td>
<td>31.54</td>
</tr>
<tr>
<td>Nearest Centroidh</td>
<td>30.86</td>
<td>33.16</td>
<td>32.43</td>
</tr>
<tr>
<td>Localized Linear Discriminant Analysisj</td>
<td>27.29</td>
<td>32.44</td>
<td>32.06</td>
</tr>
<tr>
<td>Naive Bayeld</td>
<td>29.14</td>
<td>32.19</td>
<td>31.16</td>
</tr>
<tr>
<td>Projection Pursuit LDA Treek</td>
<td>28.45</td>
<td>31.93</td>
<td>31.28</td>
</tr>
<tr>
<td>Linear Discriminant Analysisl</td>
<td>27.03</td>
<td>31.86</td>
<td>31.16</td>
</tr>
<tr>
<td>Multinomial Logistic Regressionm</td>
<td>27.02</td>
<td>31.78</td>
<td>31.14</td>
</tr>
<tr>
<td>Stump Decision Treesn</td>
<td>30.46</td>
<td>30.50</td>
<td>31.15</td>
</tr>
<tr>
<td>Artificial Neural Networko</td>
<td>27.94</td>
<td>30.50</td>
<td>30.90</td>
</tr>
<tr>
<td>Binary Decision Treesp</td>
<td>30.06</td>
<td>30.18</td>
<td>30.87</td>
</tr>
<tr>
<td>Conditional Inference Treed</td>
<td>30.03</td>
<td>29.60</td>
<td>30.55</td>
</tr>
<tr>
<td>K-Nearest Neighborf</td>
<td>26.75</td>
<td>29.11</td>
<td>28.31</td>
</tr>
<tr>
<td>C4.5 Decision Treef</td>
<td>27.43</td>
<td>28.03</td>
<td>29.07</td>
</tr>
<tr>
<td>Part Decision Treedf</td>
<td>27.16</td>
<td>28.99</td>
<td>28.60</td>
</tr>
<tr>
<td>Simple Linear Regressionw</td>
<td>27.04</td>
<td>28.79</td>
<td>27.04</td>
</tr>
<tr>
<td>Ripper Rule Learnerg</td>
<td>25.82</td>
<td>26.10</td>
<td>25.43</td>
</tr>
<tr>
<td>Random Guessingw</td>
<td>24.99</td>
<td>25.02</td>
<td>25.01</td>
</tr>
</tbody>
</table>

a Algorithm of Tibshirani et al. (31). Similar to Nearest Centroid approach, except class centroids are standardized and “shrunk” towards an overall centroid before evaluation of test cases. [pamr, pamr.train]

b Left-spherically distributed linear scores are derived from predictor variables following the dimensionality reduction rule of Laeuter et al. (59). Linear scores are used as inputs for standard linear discriminant analysis. [pred, slda]

c Predictor variables are mapped to a higher dimensional space using a specified kernel function. A linear hyperplane is identified with the largest possible margin between the two classes to be distinguished, and this “maximal margin” hyperplane is used to classify test cases. The listed accuracies were obtained
using a polynomial kernel function, with model parameters chosen by searching possible values and identifying those that minimize prediction errors on the training data. An introduction to support vector machines is provided by Byvatov and Schneider [60].

Training data are modeled as a Gaussian Process, with mean and covariance functions partly determined by parameters estimated during model training. Within this framework, class probabilities associated with each test case are estimated, as described by Williams and Barber [61]. Density estimation was performed using the radial basis kernel. [kernlab, gausspr]

Similar to Random Forest, except conditional inference trees are used as a base classifier. Listed accuracy obtained using 100 decision trees per forest [party, cforest]

Algorithm of Breiman [29]. Predictor variables and training examples are randomly selected to construct a "forest" of decision trees. Test cases are then classified by a voting procedure among all trees in the forest. Listed accuracy was obtained by growing 1000 decision trees per simulation. [randomForest, randomForest]

Implements support vector training method of Platt [62]. [RWeka, SMO]

Centroids are computed for each class using the training data. Test cases are then assigned the class label of the most similar centroid. [klaR, nm]

Implements the linear discriminant analysis approach proposed by Tutz and Binder [63] [klaR, loclda]

A probability density function is estimated for each class based upon training data. For each test case, the estimated density is used to compute the probability of class membership for each class (assuming conditional independence of class-conditional probabilities). Test cases are assigned to the class for which its class-conditional probability is greatest. [klaR, NaiveBayes]

Decision tree in which projection pursuit linear discriminant functions are used as attributes at each node [64] [classPP, PP.tree]

Linear combinations of predictor variables are constructed to maximize the ratio of variation between classes versus the variance within classes. This provides a subspace of predictors with lower dimensionality that is used for classification of test cases. [klaR, lda]

Multinomial logistic regression models based on ridge regression. See le Cessie and van Houwelingen [65] [RWeka, Logistic]

Binary decision trees are constructed using only one predictor variable. [Rweka, DecisionStamp]

A neural network with one hidden layer, 4 output nodes (one for each class) and number of input nodes equal to the number of predictor variables. During the training stage, a set of network weights that minimizes training errors is iteratively identified, which determines the contribution of each input variable to the overall network response. Basheer and Hajmeer [66] provide an introduction to this approach. [nnet, nnet]

Binary decision trees grown by recursive partitioning. Predictor variables were split to maximize information gain. [tree, tree]

Two-step algorithm of Hothorn et al. [67]. The variable most strongly associated with class labels among training examples is selected and a decision tree branch is formed through a binary split of this chosen variable. The process repeats until all predictors significantly associated with class labels (at level α) have been incorporated. Listed accuracy obtained using α = 0.30. [party, ctree]

For each test case, the k most similar observations among the training data are identified. The test case is assigned the most frequently occurring class label among the k most similar training observations. Listed accuracy obtained for k = 5. (see 68). [klaR, knn]

Decision tree induction following Hunt’s algorithm, in which trees are recursively grown by splitting attributes to maximize information gain [69]. [RWeka, J48]

Partial decision trees [70] [RWeka, PART]

Least squares multiple regression. Lifespan quartile is treated as a numeric ordinal variable and model selection is performed using the Akaike criterion. [RWeka, LinearRegression]

Test cases are classified according to a series of “if…then” rules extracted from training data using the RIPPER algorithm [71]. [RWeka, JRip]

No model building is performed and class labels are randomly assigned to test cases.