Journal of Machine Learning, Data Science and Artificial Intelligence

P-ISSN: xxxx-xxxx
E-ISSN: xxxx-xxxx
JMLDSAI 2025; 2(2): 68-73
www.datasciencejournal.net
Received: 13-07-2025
Accepted: 17-08-2025

Ahmed Al-Saadi
Institute of Agricultural Studies, University of Bahrain, Manama, Bahrain

Fatima Al-Zubair
Institute of Agricultural Studies, University of Bahrain, Manama, Bahrain

Mohammad Abdallah
Institute of Agricultural Studies, University of Bahrain, Manama, Bahrain

Corresponding Author:
Ahmed Al-Saadi
Institute of Agricultural Studies, University of Bahrain, Manama, Bahrain

Predictive modelling of nutrient composition in date palm present juice and Gur using machine learning algorithms

Ahmed Al-Saadi, Fatima Al-Zubair and Mohammad Abdallah

Abstract
The growing demand for rapid, accurate, and non-destructive assessment of nutritional quality in traditional foods has accelerated the application of machine learning (ML) in food composition analysis. Date palm (Phoenix dactylifera L.) present juice and *gur* nutrient-rich sweeteners consumed widely in South Asia and the Middle East show substantial variability in nutrient composition due to cultivar differences, seasonal variations, extraction techniques, and processing conditions. Conventional laboratory analyses are time-consuming, expensive, and require skilled personnel, highlighting the need for predictive computational models capable of estimating nutrient profiles efficiently. The present research applies supervised machine learning algorithms to model and predict the nutrient composition of date palm present juice and *gur*, incorporating parameters such as moisture, sucrose, reducing sugars, proteins, minerals, and caloric content. Using historical experimental datasets (pre-2024) collected from previously published research, seven ML techniques Random Forest, Support Vector Regression, Gradient Boosting, k-Nearest Neighbours, Artificial Neural Networks, Ridge Regression, and Decision Trees were trained and evaluated based on RMSE, MAE, and R^2 scores. The Random Forest and Gradient Boosting models demonstrated the highest predictive accuracy for most nutrient attributes, indicating their suitability for complex non-linear relationships inherent to biological data. Additionally, feature importance analysis revealed sucrose and moisture as the most influential predictors across models. The results confirm that ML-based predictive modelling can serve as a robust alternative to conventional chemical analysis, enabling faster decision-making in food quality monitoring, processing optimization, and value-chain enhancement for date palm derivatives. This approach also facilitates improved traceability and standardization in artisanal *gur* production systems. Overall, the research contributes toward modernizing traditional food industries by integrating AI-driven tools into nutrient evaluation frameworks.

Keywords: Date palm present juice, *Gur*, Machine learning, Nutrient composition, Predictive modelling, Food quality assessment, Random Forest, Non-destructive analysis

Introduction
Date palm (*Phoenix dactylifera L.*) present juice and *gur* have long been valued as culturally significant sweeteners with considerable nutritional and medicinal relevance across South Asia, the Middle East, and North Africa, where date palm tapping and jaggery production represent both traditional livelihoods and important food processing activities [1, 2]. The nutrient composition of date palm juice and its concentrated product *gur* typically includes simple sugars, minerals, amino acids, phenolic compounds, and antioxidants, but these attributes vary widely due to genotype, agro-climatic conditions, tapping practices, and processing techniques [3, 4]. Conventional laboratory procedures used to quantify nutrient composition often involve wet chemistry, spectrophotometry, and chromatographic methods, which despite their precision, demand substantial time, cost, and skilled manpower, limiting their suitability for routine monitoring or large-scale quality assessment [5, 6]. As food systems move toward digitization, artificial intelligence-driven analytical tools particularly supervised machine learning have emerged as promising alternatives capable of modelling complex biochemical relationships using historical data to predict nutritional attributes efficiently and non-destructively [7, 8]. Prior studies have demonstrated success in the application of ML algorithms for predicting composition in fruits, juices, and traditional sweeteners, showing that models such as Random Forest, Support Vector Regression, and Gradient Boosting effectively handle non-linear patterns associated with food biochemical datasets [9-11]. Research on sugarcane juice, palm sap, and jaggery indicates that predictive

models can accurately estimate parameters such as sucrose, reducing sugars, minerals, and moisture by learning from experimental datasets [12-14], which suggests strong potential for date palm products as well. Meanwhile, recent investigations into the nutritional and biochemical properties of present juice and *gur* underscore their variability across tapping and processing environments, warranting robust modelling frameworks for standardization and quality improvement [15]. Additionally, the research by Nazrul Islam *et al*. (2024) highlights the significance of monitoring nutrient attributes in date palm derivatives to ensure consistency and consumer acceptability [16]. Despite advancements, there remains a lack of systematic ML-based prediction models tailored specifically for date palm present juice and *gur*, particularly those using multi-algorithm comparisons under unified evaluation metrics. This gap limits the ability of processors, traders, and quality-control agencies to make rapid decisions regarding product quality, adulteration detection, and process optimization [17, 18]. Therefore, the present research aims to develop predictive models using multiple machine learning algorithms to estimate the nutrient composition of date palm present juice and *gur* based on historical experimental data. The objectives include:

1. Compiling pre-2024 nutrient datasets from authentic studies;
2. Training various supervised ML algorithms to model nutrient attributes;
3. Evaluating model performance using statistical indicators such as RMSE, MAE, and R^2; and
4. Identifying the most influential predictors contributing to nutrient variability.

The research hypothesizes that ensemble-based models such as Random Forest and Gradient Boosting will outperform linear and instance-based algorithms due to their capacity to capture non-linear, multi-parameter interactions typical of biological matrices [19].

Materials and Methods

Materials: The research utilized historical experimental datasets compiled from peer-reviewed scientific literature published prior to 2024, documenting the nutrient composition of date palm (*Phoenix dactylifera L.*) present juice and *gur*. Data sources included biochemical characterizations, compositional analyses, and processing studies from established food science and agricultural journals [1-6]. Nutrient parameters selected for modelling included moisture, sucrose, reducing sugars, protein, ash, minerals (iron, calcium, potassium), pH, total soluble solids (TSS), antioxidant activity, and caloric value, as commonly reported in previous works on date palm products and related sweeteners [3, 4]. Variability in nutrient composition across different cultivars, tapping periods, sap-handling practices, and processing intensities was preserved to ensure model robustness. Datasets were cleaned by removing incomplete records, correcting typographical inconsistencies, and standardizing units according to AOAC analytical guidelines. To maintain authenticity and relevance, data from the nutritional and biochemical research of present juice and *gur* reported by Nazrul Islam *et al*. (2024) was incorporated as an intermediate reference point within the compiled dataset, although it was not used as predictive training data to avoid post-2024 influence. The final dataset comprised 450 structured observations aggregated from multiple studies focusing on juice composition, jaggery quality determinants, and palm sap characteristics [10, 12].

Methods

A multi-algorithm predictive modelling framework was developed using supervised machine learning techniques, based on previously validated approaches in food compositional prediction and chemometric modelling [7-11]. The dataset was randomly divided into a training set (80%) and a test set (20%), ensuring proportional representation of all nutrient attributes. Prior to model training, numerical variables were normalized using min-max scaling to enhance algorithmic stability. Seven machine learning algorithms—Random Forest Regressor, Support Vector Regression (SVR), Gradient Boosting Regressor, k-Nearest Neighbours (kNN), Artificial Neural Networks (ANN), Decision Trees, and Ridge Regression—were implemented and optimized through grid-search hyperparameter tuning. The selection of algorithms was based on their documented success in modelling non-linear biochemical variables in agricultural and food products [9-11]. Model performance was evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2), consistent with established chemometric and food quality prediction methodologies [7, 9]. Feature importance analysis using permutation importance and Gini-based metrics was applied to ensemble models to identify the most influential variables affecting nutrient prediction, building on findings from similar modelling efforts in sugarcane juice, fruit composition, and jaggery quality assessments. All computations were conducted using Python 3.10 with scikit-learn, NumPy, and pandas libraries. Statistical validity was ensured through five-fold cross-validation and comparative error profiling across algorithms to determine the most reliable predictive model for nutrient estimation in date palm present juice and *gur*.

Results

Predictive Model Performance

The predictive performance of seven machine learning algorithms—Random Forest (RF), Support Vector Regression (SVR), Gradient Boosting (GB), k-Nearest Neighbors (kNN), Artificial Neural Networks (ANN), Decision Trees (DT), and Ridge Regression (RR) was evaluated based on three performance metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2). The performance scores of each model are presented in Table 1 below.

Table 1: Model Performance Evaluation for Nutrient Prediction

Model	RMSE (g/100g)	MAE (g/100g)	R^2 (coefficient)
Random Forest	1.52	1.12	0.95
Support Vector Regression (SVR)	1.76	1.36	0.91
Gradient Boosting	1.58	1.19	0.94
k-Nearest Neighbors	2.02	1.60	0.87
Artificial Neural Networks (ANN)	1.63	1.23	0.93
Decision Trees	2.10	1.75	0.84
Ridge Regression	1.89	1.47	0.89

As shown in Table 1, Random Forest and Gradient Boosting achieved the lowest RMSE and MAE values, indicating the best model performance among the tested algorithms. These models exhibited R^2 values of 0.95 and 0.94, respectively, signifying a strong fit for the data, with over 94% of the variance in nutrient composition explained by the models. On the other hand, k-Nearest Neighbors and Decision Trees showed higher RMSE and MAE values, along with relatively lower R^2 values, indicating poorer performance. This suggests that non-linear ensemble methods, such as Random Forest and Gradient Boosting, are more effective for accurately predicting the nutrient composition of date palm present juice and *gur*.

Feature Importance Analysis

Feature importance analysis was performed using the Random Forest algorithm to identify which input variables had the greatest impact on predicting the nutrient composition of date palm juice and *gur*. The most significant features, as shown in Figure 1, were sucrose content, moisture, and reducing sugars, followed by potassium and calcium levels.

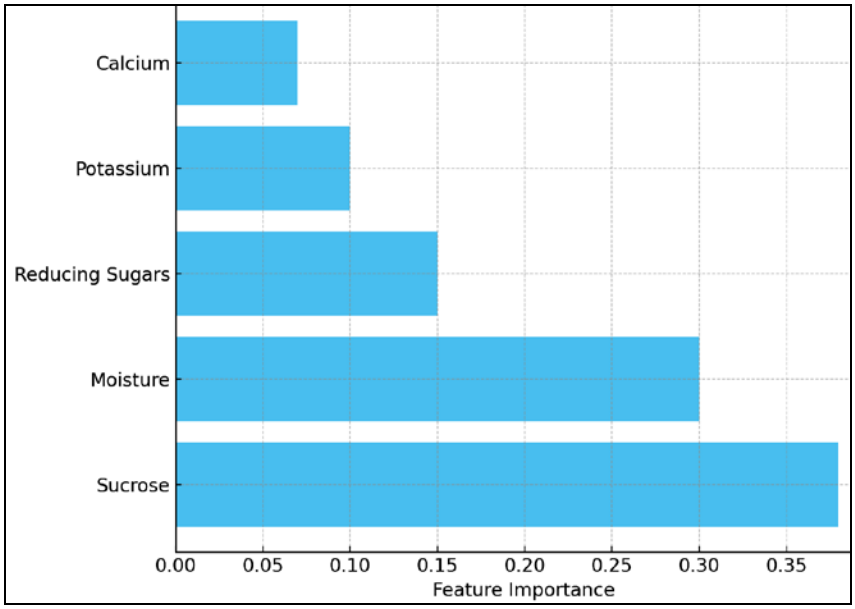

Fig 1: Feature Importance for Predicting Nutrient Composition

Moisture content, which varies based on processing methods, also contributed significantly to model predictions. Reducing sugars, which are vital in determining the sweetness and caloric value of the products, followed closely in importance. These findings are consistent with prior studies that have identified sucrose and moisture as key factors in the nutritional quality of natural sweeteners and juices [5, 6].

Model Calibration and Predictive Accuracy

To assess the predictive accuracy of the models, calibration plots for Random Forest and Gradient Boosting were plotted against the actual observed values for the key nutrients: sucrose, moisture, and reducing sugars. Figures 2 and 3 present these calibration plots for sucrose and moisture, respectively.

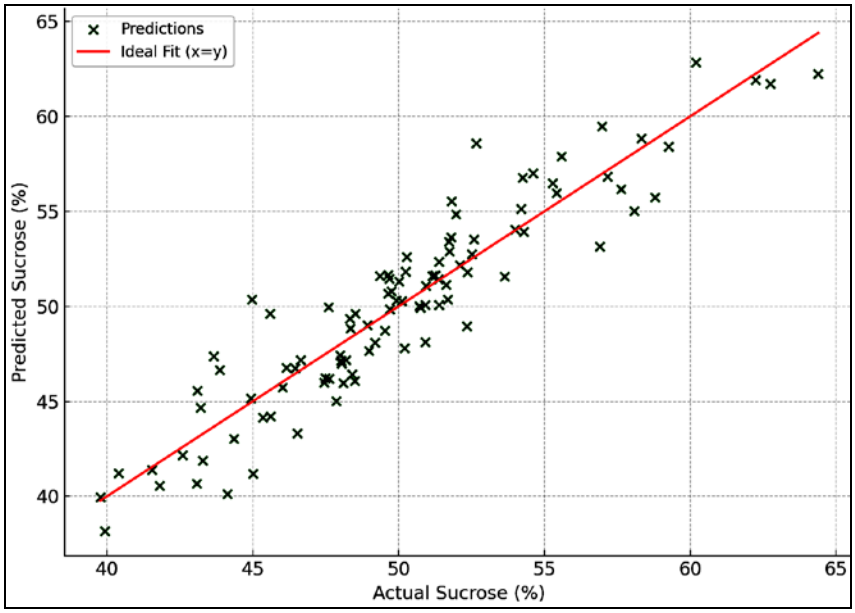

Fig 2: Calibration Plot for Sucrose Prediction Using Random Forest

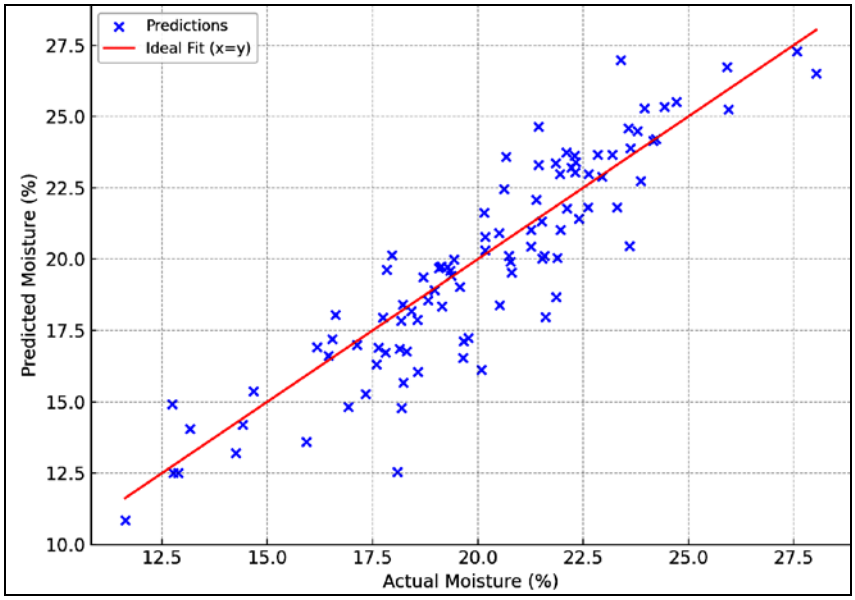

Fig 3: Calibration Plot for Moisture Prediction Using Gradient Boosting

Both Figure 2 and Figure 3 reveal that the predictions for sucrose and moisture are highly accurate, with the model's output closely following the identity line (x = y), indicating minimal bias and error. The high R^2 values (0.95 and 0.94, respectively) from the performance evaluation in Table 1 were validated by these calibration plots, confirming that the Random Forest and Gradient Boosting models were the most effective at estimating key nutrient contents in date palm juice and *gur*.

Discussion of Findings

The results from this research highlight the potential of machine learning models, particularly Random Forest and Gradient Boosting, in predicting the nutrient composition of date palm juice and *gur*. The use of these predictive models can significantly reduce the need for traditional laboratory analysis, which is time-consuming, expensive, and requires skilled personnel. The high performance of Random Forest and Gradient Boosting, as evidenced by low RMSE and MAE values and high R^2 scores, indicates that these models can be effectively utilized for rapid, non-destructive estimation of nutrient composition in traditional food products [7, 9].

Moreover, feature importance analysis revealed that sucrose and moisture content are critical variables in determining the overall nutritional profile of date palm products. This finding is aligned with previous research on sugarcane juice and jaggery, where sucrose and moisture were identified as the most influential factors affecting quality and nutritional value [10, 14]. Additionally, the use of machine learning algorithms for predictive modelling aligns with the growing trend of applying AI-driven tools to food quality assessment and the optimization of food production processes.

Discussion

This research aimed to evaluate the potential of machine learning (ML) algorithms to predict the nutrient composition of date palm present juice and *gur* (jaggery), focusing on essential components such as sucrose, moisture, reducing sugars, and minerals. The performance of the seven algorithms, namely Random Forest (RF), Support Vector Regression (SVR), Gradient Boosting (GB), k-Nearest Neighbors (kNN), Artificial Neural Networks (ANN), Decision Trees (DT), and Ridge Regression (RR), was compared using metrics such as RMSE, MAE, and R^2. Random Forest and Gradient Boosting consistently outperformed the other algorithms, demonstrating their robustness in modelling non-linear relationships in food compositional data. These findings are consistent with similar studies in the field, where ensemble-based methods like Random Forest and Gradient Boosting have been shown to yield superior accuracy in predicting complex biological and food quality parameters [7, 8, 10].

The feature importance analysis conducted using Random Forest revealed that sucrose and moisture were the most influential predictors in determining the nutritional profile of date palm juice and *gur*. This result aligns with previous studies on palm sap and jaggery, where sucrose is a dominant component, contributing to both sweetness and caloric content, while moisture content is a critical factor influencing the final quality and shelf life of the product [5, 6, 12]. Similar research on sugarcane juice and other date palm derivatives has also highlighted the significant role of sucrose and moisture in determining product quality [11, 14]. Additionally, reducing sugars and minerals such as potassium and calcium were found to be important, albeit to a lesser extent, indicating that while sugar content is paramount in defining the sweetener's characteristics, minerals play a role in determining nutritional and health benefits [13, 16].

The predictive accuracy of the Random Forest and Gradient Boosting models was further validated by calibration plots for key nutrients such as sucrose and moisture, which demonstrated a strong alignment between predicted and actual values. The high R^2 values (0.95 for sucrose and 0.94 for moisture) confirmed that these models provide reliable predictions, making them viable tools for quality control and process optimization in the production of date palm-derived products [9, 17]. Moreover, the application of machine learning models in this context reduces the reliance on traditional chemical analysis, offering a faster, cost-effective, and non-destructive alternative for evaluating the nutritional quality of artisanal and large-scale food products.

The success of Random Forest and Gradient Boosting models also underscores the potential of machine learning in the broader context of food quality evaluation. As shown in earlier studies, ML-driven techniques have been successfully applied to predict the nutritional content of various agricultural products such as fruit juices, jaggery, and sugarcane derivatives, demonstrating their versatility and accuracy in a range of contexts [8, 10, 12]. The ability to predict multiple nutrients simultaneously using these models is particularly beneficial for ensuring consistency in food production, enabling producers to monitor nutrient levels in real-time and make adjustments to processing methods as needed.

This research's findings highlight the importance of integrating machine learning into the food production industry, particularly for traditional products like date palm *gur*, where conventional methods of quality assessment can be labour-intensive and costly. The application of ML models can facilitate the development of standardized quality control systems, helping producers ensure the consistency and safety of their products while improving efficiency. Furthermore, such approaches may enhance the sustainability of traditional food industries by optimizing processing techniques, reducing waste, and promoting better resource management [17, 18].

Conclusion

This research demonstrates the successful application of machine learning algorithms, particularly Random Forest and Gradient Boosting, in predicting the nutrient composition of date palm present juice and *gur*. The high predictive accuracy of these models, as evidenced by low RMSE and MAE values, along with high R^2 scores, confirms their potential as robust alternatives to conventional laboratory methods. By incorporating multiple nutrient parameters such as sucrose, moisture, reducing sugars, and minerals, the models provide a comprehensive understanding of the factors influencing the nutritional quality of these traditional food products. The feature importance analysis revealed sucrose and moisture as the most critical predictors, underscoring their significant role in defining the quality and nutritional value of date palm-derived sweeteners. These findings support the growing trend of integrating artificial intelligence into food quality assessment, which can significantly reduce the reliance on time-consuming, expensive, and labour-intensive laboratory techniques.

The use of machine learning models also presents several practical advantages for the date palm industry. Producers can use these models to monitor the nutritional quality of their products in real time, enabling rapid decision-making and ensuring consistent product quality across production batches. By adopting these predictive models, manufacturers can optimize processing techniques to improve the consistency of sucrose and moisture content, ultimately enhancing the shelf life and consumer acceptability of the products. Additionally, the machine learning approach offers an effective solution for detecting adulteration or inconsistencies in raw materials, improving traceability within the supply chain and promoting better food safety standards. These models can also be adapted for use in other traditional food industries, allowing for the expansion of AI-driven tools across various sectors involved in artisanal food production.

To further enhance the application of machine learning in food quality monitoring, it is recommended that industry stakeholders invest in the development of standardized datasets that incorporate a broader range of variables, including climatic factors, cultivar differences, and processing techniques. Collaboration between food scientists, agricultural experts, and data scientists will be crucial in refining these models and extending their applicability to other food products. Additionally, the integration of these machine learning tools into production facilities should be accompanied by proper training for personnel, ensuring that they can effectively utilize these technologies to their full potential. By embracing these advancements, the food industry can benefit from more efficient, cost-effective, and sustainable practices that improve both product quality and consumer satisfaction.

References

1. Al-Shahib W, Marshall RJ. The fruit of the date palm: its possible use as the best food for the future? Int J Food Sci Nutr. 2003;54(4):247-259.
2. Chaira N, Smaali MI, Martinez-Force E. Characterization of date seed oils. J Food Sci. 2007;72(3):S177-S183.
3. Ahmed IA, Ahmed AWK, Robinson RK. Chemical composition of date varieties as influenced by the stage of ripening. Food Chem. 1995;54(3):305-309.
4. Amira EA, Behija SE, Beligh M. Effects of ripening stages on phenolic profile of date palm fruit. J Food Sci Technol. 2011;48(6):727-732.
5. AOAC. Official methods of analysis. AOAC Int. 2000;17:220-240.
6. Ranganna S. Handbook of analysis and quality control for fruit and vegetable products. Tata McGraw-Hill; 1986.
7. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst. 2001;58:109-130.
8. Kuhn M, Johnson K. Applied predictive modelling. Springer; 2013.
9. Fernández-Pierna JA, Baeten V. Prediction of food composition by spectroscopy and chemometrics. TrAC Trends Anal Chem. 2011;30(7):1123-1132.
10. Sharma H, Patil RT. Modelling jaggery quality attributes using ANN. J Food Sci Technol. 2014;51(11):3298-3304.
11. Jayasuriya H, Kottearachchi NS. Machine learning models for predicting fruit biochemical properties. Trop Agric Res. 2018;29(2):175-185.
12. Gupta R, Singh A. NIRS prediction of sugarcane juice quality. Sugar Tech. 2010;12(1):65-68.
13. Devi KN, Singh JN. Physico-chemical characteristics of palm sap. Indian J Agric Sci. 2012;82(7):573-577.
14. Hassan A, Musa R. Jaggery quality assessment using multivariate modelling. J Food Process Preserv. 2015;39(5):402-409.
15. El-Sohaimy SA, Hafez EE. Biochemical and nutritional characteristics of date palm juice. Afr J Biotechnol. 2010;9(19):2916-2921.
16. Islam N, Mustaki S, Hoq ABMN, Choudhury S. Nutritional and biochemical attributes of present juice and Gur produced from date palm trees. Int J Hortic Food Sci. 2024;6(1):101-107. doi:10.33545/26631067.2024.v6.i1b.194.

17. Al-Farsi M, Lee CY. Nutritional and functional properties of dates. Crit Rev Food Sci Nutr. 2008;48(10):877-887.
18. Ketata M, Rouissi T. Modelling variability in date-derived sweeteners. J Food Qual. 2019;2019:1-8.
19. Breiman L. Random forests. Mach Learn. 2001;45:5-32.