Ensemble Machine Learning to Predict and Feature Engineering to Identify Factors for Academic-Success

Syed Affan Daimi[1]*, Asma Iqbal[2]

[1] Muffakham Jah College of Engineering & Technology, India
[2] Deccan college of Engineering & Technology, India.
saffand03@gmail.com

Abstract. This paper presents academic-success, a machine learning predictor system designed to forecast student dropout rates or academic success and offer tailored interventions and support. The work encompasses the design of a predictive model and the provision of resources to integrate the model into existing student support systems within educational institutions.

The proposed predictive model is implemented and incorporating feature engineering, Importance analysis is carried out. The "enrolled class" is included for training and validation of the predictive model and also for the Importance analysis. The experiments are done using the ensemble learning and Support Vector Classifier (SVC). According to the experimental results for Area Under Curve in Receiver Operating Characteristics (ROC) curve, inclusion of the "Enrolled class" gave improved results for both ensemble learning (0.91) as well as SVC (0.86). Feature importance analysis using Leave One Feature Out (LOFO) and SHapley Additive exPlanations (SHAP) gave interesting results on the significance of education/occupation of the mother vs the father's, importance of a steady approach towards course enrollment and the role of scholarships. This gives an added impetus to focus on women's education and empowerment to improve the society through the next generation.

Keywords: Education, Random Forest, SVM, Hugging face, MLOps

1 Introduction

The need to reduce the dropout rate is a global challenge and is a greater challenge in the developing countries. Constructive steps taken to help students complete their education shall not only be beneficial to the student's future success in life [1,2], but will also be an investment for the development of a country. In recent years, predicting student dropout rates and identifying factors contributing to academic success has gained significant attention in the field of education as it helps early intervention and provide targeted support, ultimately improving student outcomes. Machine learning techniques have emerged as powerful tools for addressing this challenge, leveraging large datasets to develop predictive models.

C. Kiran Mai et al. (eds.), *Proceedings of the Fourth International Conference on Advances in Computer Engineering and Communication Systems (ICACECS 2023)*, Atlantis Highlights in Computer Sciences 18,
https://doi.org/10.2991/978-94-6463-314-6_28

This paper presents academic-success, a machine learning predictor system designed to forecast student dropout rates or academic success and provide tailored interventions and support. The objective of this work is to develop a robust predictive model that integrates seamlessly into existing student support systems by leveraging machine learning algorithms. Academic-success aims to assist educational stakeholders in identifying students who are at risk of dropping out.

In this paper, we present the methodology and development process of the academic-success system. The dataset used is obtained from an institution of higher education of students enrolled in undergraduate studies in the fields of technology, nursing, journalism, management, design and education to name a few. We present a brief study of the literature and work done in this area in section II. Section III comprises of the methodology followed for this work, section IV contains the experimental results followed by conclusions and future possibilities in section V.

2 Background

The application of machine learning techniques to predict the dropout or quit vs completion or retention is quite powerful and sturdy method to obtain insightful results. The focus of the considerable work done in this area is on e-learning, classroom studies, on data from developed countries, university level education and using data from higher secondary and first year courses. A survey on machine learning techniques and its challenges was presented by Mduma N. et al. [3]. Márquez-Vera et al. [4] used genetic programming in conjunction with different machine learning algorithms. The metrics- True Positives (TP), True Negatives (TN) and accuracy were considered to identify the best model. Another scheme put forth by Knowles [5], used a predictive machine learning model to find student dropout risk. Here, the metric used to identify the best model is the ROC curve that is traditionally most often used.

Dalipi F. et al. [6] presented a dropout prediction machine learning based method for MOOCs. The factors were segregated into student related and MOOCs methodology related issues. The frequency of the various machine learning and deep learning algorithms used in earlier studies was also included. The ensemble learning algorithms haven't been explored for this. The authors have consolidated the earlier works, and put forth suggestions both for students and MOOCs methodology so as to diminish student dropout rate. Li, H. et al. [7] worked on the hybrid mode of teaching that was partially online along with the traditional classroom teaching. Several machine learning models were used like Support Vector Regression, Naive Bayes and K-Nearest Neighbour etc. to study the expected course grades of students. Support Vector Machine (SVM) gave the best performance among all the designed models.

A few studies were based on collecting data from students. In the work done by Martinho et al. [8] a questionnaire was used to collect data and then select relevant features. Some of the features like marital status, family finance etc. were indicative of the students' family status while others like school of origin, conveyance etc. were reflective of life conditions of the student. Fuzzy-ARTMAP neural network was used as it enables continuous learning and the metric used system was its accuracy measure. Another approach for analysing the dropout rate and helping improve it was presented by Kadar M. et al. [9]. The required data was collected from devices like web-cam, eye tracker etc in the smart classrooms. Feature selection [10] techniques were used to short-list important features. Additional data on sign language for the deaf was also collected. The study in conclusion gave constructive inputs to the professors, improved software accessibility and included sign translators

3 Methodology

The dataset used in this paper has a multi-class target variable i.e. {'enrolled', 'dropout', 'graduated'}. As the target variables of interest are {'dropout', 'graduated'}, the data points with target variable {'enrolled'} are neglected.

Considering the high dimensionality of the data, we experiment with State Vector Machine [11] based algorithms and ensemble-based algorithms, starting with a baseline model using a Random Forest Classifier [12,13], followed by further exploration of data to achieve improved performance. Additionally, we address the handling of the "Enrolled" class of the dataset, which was initially deemed irrelevant but later recognized as a valuable source of information for predictions. The inclusion of the “Enrolled” data points for test and validation has helped address the problem of the imbalance in the data. Furthermore, we explore feature engineering techniques and conduct feature importance analysis to identify the key factors influencing student outcomes.

3.1 Baseline Model Development

To develop the baseline model, we explored the usage of two popular machine learning algorithms: Support Vector Classifier (SVC) and Random Forest. These algorithms were chosen for their ability to handle complex datasets and provide competitive results in various classification tasks.

The Support Vector Classifier (SVC) is a supervised learning algorithm that is effective in separating data points into different classes by constructing hyperplanes in a high-dimensional feature space. It aims to find the optimal decision boundary that maximizes the margin between classes while minimizing classification errors. It has emerged as a popular choice when dealing with high dimensional data.

On the other hand, Random Forest is a method of ensemble learning that merges the outputs of multiple decision trees to make predictions. The individual decision trees are constructed by randomly selecting subsets of features and data samples from the training set. During training, the trees independently learn different aspects of the data and make predictions. By collectively considering the predictions of all individual trees, typically through majority voting for classification, the random forest model determines its ultimate prediction.

Through extensive experimentation and evaluation, we found that both algorithms performed equally well in predicting student dropout rates or academic success. The competitive results obtained from both models demonstrated their capability to handle the complexity of the dataset effectively.

In a nutshell, the baseline model development involved exploring the usage of SVC and Random Forest algorithms, both of which demonstrated equally good performance in predicting student dropout rates or academic success. These algorithms were selected as the foundation for further analysis and improvement in the subsequent stages of the project.

3.2 Handling the 'Enrolled' Class

The "Enrolled" class represents students who are currently enrolled and have yet to graduate or drop out. Despite initial doubts regarding its relevance, further analysis revealed its importance for predicting future outcomes of these students. Given the relatively small dataset size, it is advantageous to make use of 100% of the available data, including the "Enrolled" class. To handle the "Enrolled" class effectively, we focused on utilizing the Random Forest Classifier, which outperformed the Support Vector Classifier (SVC) in our experiments. Table 1 showcases the performance metrics obtained from the baseline models using Random Forests and SVM algorithms. It presents important evaluation measures, such as F1 score, accuracy, precision, and recall. These metrics provide insights into the initial performance of the models before any modifications or enhancements were implemented. Thus, we made the decision to retain only the Random Forest model for subsequent analysis.

Table 1. Metrics from the baseline model (SVC & Random Forest)

Model	**F1 Score**	**Accuracy**	**Precision**	**Recall**
Random Forest	0.75	0.77	0.75	0.77
SVC	0.71	0.73	0.71	0.73

To incorporate the "Enrolled" class into the model, we adopted a two-step approach. Firstly, we trained a Random Forest model exclusively on the "Graduated" and "Dropout" classes, which were the primary classes of interest for prediction. This model served as the foundation for predicting outcomes for the "Enrolled" class.

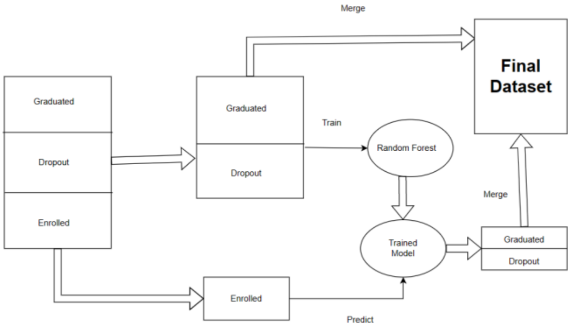

Fig. 1. Method to handle the "Enrolled" data points.

Next, we employed the trained Random Forest model to predict the outcomes for the "Enrolled" students in our original dataset. By replacing the original "Enrolled" labels with the predicted outcomes generated by the Random Forest model, we constructed a modified dataset. This modified dataset contained predictions for only two classes while preserving the original dataset's features. The detailed steps involved are shown in figure 1.

3.3 Feature engineering & Importance analysis

3.3.1 Feature Engineering [14]:

The dataset was subjected to feature engineering techniques which initially had thirty (30) features as listed in Table 2. During this process, the significance of various features was evaluated to determine their impact on the prediction task. Notably, the current marks of the students emerged as a highly important feature, indicating its strong correlation with student outcomes.

Due to this, the importance of the "Enrolled" category became apparent. Retaining this class in the analysis proved crucial, as it contributed essential information about students currently enrolled, distinct from those who have graduated or dropped out. This finding underscores the relevance and value of considering the "Enrolled" category when predicting student outcomes.

Table 2. Set of Features

Marital Status	**Age at Enrollment**	**Father's Qualifications**	**International**	**Mother's Qualifications**
Application Mode	Curricular units 1st (credited)	Displaced	Curricular units 1st (enrolled)	Inflation rate
Application Order	Curricular units 1st (evaluations)	Educational Special Needs	Curricular units 1st (approved)	Mother's occupation
Course	Curricular unit's 1st (grade)	Debtor	Curricular units 1st (without evaluations)	Unemployment rate
Daytime/ Evening Attendance	Curricular units 2nd (credited)	Tuition Fees up to Date	Curricular units 2nd (enrolled)	Scholarship Holder
Previous Qualifications	Curricular unit's 2nd (evaluations)	Gender	Curricular units 2nd (approved)	Father's occupation
Nationality	Curricular units 2nd (grade)	GDP	Curricular units 2nd (without evaluations)	

3.3.2 Feature Importance:

To quantify the relative importance of features, the SHAP (SHapley Additive exPlanations) [15-16] and LOFO (Leave One Feature Out) [17] importance method were employed. LOFO involved systematically excluding each feature and measuring its influence on the model's performance. SHAP (SHapley Additive exPlanations) is a method that uses game theory to interpret the results generated by machine learning models. SHAP offers localized explanations specific to individual instances, allowing users to understand how each feature affects a particular prediction. It is applicable to various machine learning models and data types, making it a versatile tool for interpreting and enhancing the transparency of machine learning systems.

4 Experimental Results:

4.1 Ensemble model predictions

We now display the evaluations of our models and the effects of our improvements. The following metrics are used to ensure robust evaluation of our work:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \tag{1}$$

$$Precision = TP / (TP + FP) \tag{2}$$

$$Recall = TP / (TP + FN) \tag{3}$$

$$F1 - score = 2 * (Precision * Recall) / (Precision + Recall) \tag{4}$$

$$ROC\ AUC\ Score = Area\ under\ ROC\ curve = TPR / (TPR + FPR) \tag{5}$$

Where True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN), True Positive Rate (TPR), False Positive Rate (FPR).

Table 3 displays the performance metrics of the models after excluding the 'Enrolled' class from the dataset and its ROC curve in figure 2.

Table 3. Performance metrics ignoring "Enrolled" students

Model	F1 Score	Accuracy	Precision	ROC
Random Forest	0.90	0.90	0.90	0.88
SVC	0.89	0.90	0.90	0.87

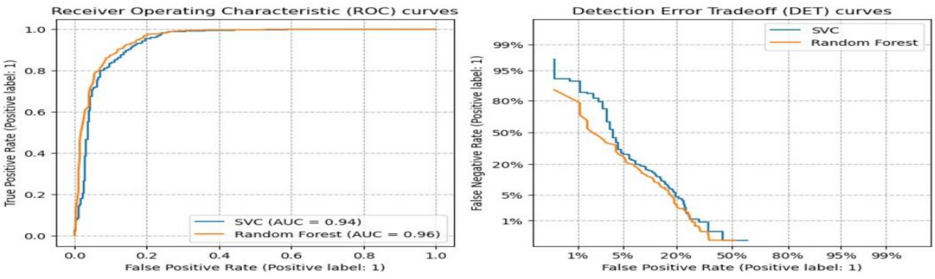

Fig. 2. ROC & DET curves for random forest & SVC models.

Table 4. Performance metrics after converting and integrating the "Enrolled" students

Model	F1 Score	Accuracy	Precision	ROC
Random Forest	0.91	0.92	0.92	0.91
SVC	0.88	0.88	0.90	0.86

Table 4 illustrates the performance metrics obtained after converting the 'Enrolled' class in the dataset and its corresponding ROC curve in figure 3.

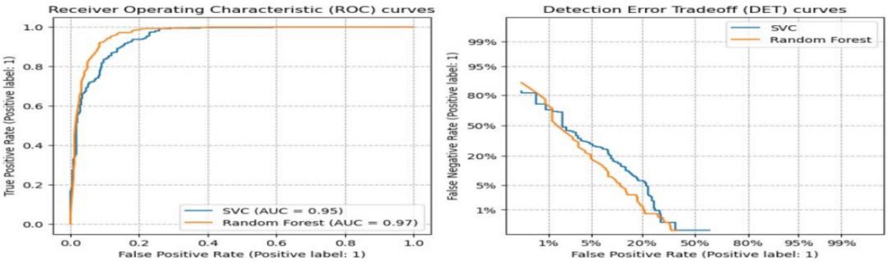

Fig. 3. ROC & DET curves for random forest & SVC models including the "Enrolled class"

The results in table 3 and table 4 indicate the efficacy of including the data points with "Enrolled" as the target variable for training and validation in case of the Random Forest model.

4.2 Inferences drawn from Feature importance analysis

The importance of features as obtained from LOFO is shown in Figure 4.

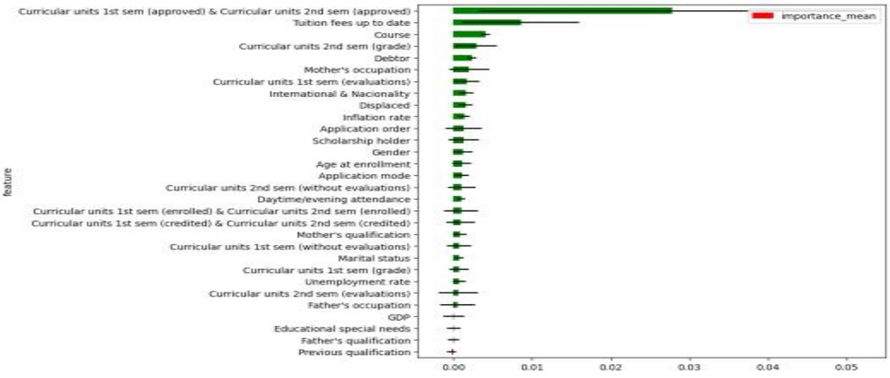

Fig. 4. Feature Importance obtained from LOFO

The analysis revealed that current marks significantly contribute to predicting student success or dropout rates. The number of courses enrolled, approved and evaluated are important factors. An interesting observation is the greater influence of the mother's education and occupation when compared to the father.

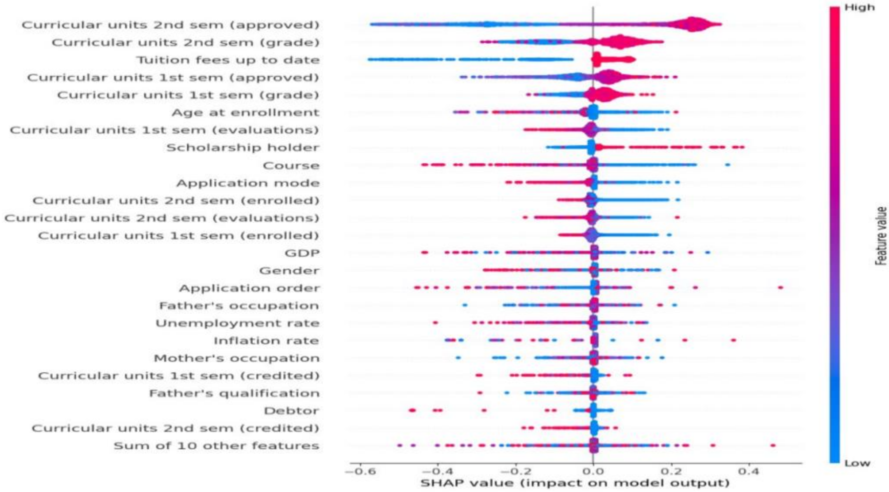

Fig. 5. Feature Analysis using SHAP

SHAP was further done to get a better insight and the output from this are shown in figure 5. The values on the right are of the graduated students and left indicates the dropouts. By prioritizing the importance of current marks and considering the unique insights provided by the "Enrolled" category, educational institutions can formulate tailored interventions and support mechanisms to cater to the unique needs of students at-risk.

5 Model Deployment and Integration

This section highlights the successful deployment of the model on the internet, integration with a user-friendly Gradio-based UI [18], the option for model export and deployment, and the importance of hyperparameter selection for achieving optimal performance.

5.1 Hugging Face [19] Deployment

The final model was deployed on the Hugging Face platform, providing a convenient and accessible solution for predicting student dropout rates and academic success. Hugging Face offers a user-friendly interface, facilitating model utilization and inference.

5.2 User Interface Using Gradio

To enhance the usability and interactivity of the deployed model, a user interface was created using Gradio. The Gradio library allowed for the development of a functional UI, enabling users to perform real-time inference using a dedicated 16GB CPU dual-core virtual machine. The interface provides a seamless experience for inputting relevant data and obtaining predictions from the academic-success model.

5.3 Model Export and Deployment

In addition to using the deployed model directly through the Hugging Face platform, users have the flexibility to export and deploy the model for their specific purposes. This enables integration into various educational institutions' existing systems or workflows. The exported model can be utilized in different environments, empowering institutions to leverage the academic-success model in their unique settings.

5.4 Hyperparameters and Training Details

The deployed model's performance is attributed to the carefully chosen hyperparameters used during the training process. Detailed information regarding the hyperparameters employed, including their specific values, is provided in the model documentation available on the Hugging Face repository (https://huggingface.co/sulpha/student_academic_success). These hyperparameters were fine-tuned to optimize the model's predictive capabilities for student dropout rates and academic success.

6 Conclusion and future Scope

6.1 Conclusion

In this study, we developed academic-success, a machine learning predictor system for forecasting student dropout rates or academic success. Through extensive experimentation and analysis, we have achieved significant advancements in accurately predicting student outcomes. The key findings and outcomes of our work are encapsulated below:

Model Development: We initially employed a Random Forest Classifier as the first model, which demonstrated promising results with 75% accuracy. Subsequently, by leveraging tree-based models, we improved the model's performance, achieving a roc_auc score of 92%.

Handling the "Enrolled" Class: Our analysis revealed the importance of retaining the "Enrolled" class, that gives insights into students currently enrolled but not yet graduated or dropped out. By creating separate models for the "Graduated" and "Dropout" classes and leveraging the first model's predictions, we enhanced the overall accuracy and predictive capabilities.

Feature Engineering and Importance Analysis: We performed feature engineering for feature importance analysis, enabling us to identify the key factors influencing a student's academic-success.

6.2 Future Work

While our current work has provided valuable insights and achieved notable results, there are several avenues for future exploration and improvement:

Mitigating Data Leakage: Future experimentation should focus on preventing data leakage for data integrity. This involves careful consideration of the data splitting strategy, feature selection, and model training techniques.

Hyperparameter Tuning [20]: The choice of the first model as a hyperparameter and experimentation with different algorithms and ensemble techniques could enhance the model's predictive performance.

Pipelining and Integration: Future work can focus on developing a robust pipeline for deploying the academic-success predictor. This would involve streamlining data preprocessing, model training, and deployment processes to create an end-to-end solution that is easily adaptable to various institutions.

By addressing these areas, future research can further enhance the accuracy, reliability, and practicality of the academic-success predictor system, enabling educational institutions to identify vulnerable students and provide targeted interventions and assistance.

References

1. Jadrić, M., Garača, Ž., & Čukušić, M. (2010). Student dropout analysis with application of data mining methods. Management: journal of contemporary management issues, 15(1), 31-46.
2. Rumberger, R. W. (1987). High school dropouts: A review of issues and evidence. Review of educational research, 57(2), 101-121.
3. Mduma, N., Kalegele, K., & Machuve, D. (2019). A survey of machine learning approaches and techniques for student dropout prediction.
4. Márquez-Vera, C., Cano, A., Romero, C., & Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. Applied intelligence, 38, 315-330.
5. Knowles, J. E. (2015). Of Needles and Haystacks: Building an Accurate Statewide Dropout Early Warning System in Wisconsin. Journal of Educational Data Mining, 7(3), 18-67.
6. Dalipi, F., Imran, A. S., & Kastrati, Z. (2018, April). MOOC dropout prediction using machine learning techniques: Review and research challenges. In 2018 IEEE global engineering education conference (EDUCON) (pp. 1007-1014). IEEE.
7. Li, H., Lynch, C. F., & Barnes, T. (2018). Early prediction of course grades: models and feature selection. arXiv preprint arXiv:1812.00843.
8. Martinho, V. R. D. C., Nunes, C., & Minussi, C. R. (2013, November). An intelligent system for prediction of school dropout risk group in higher education classroom based on artificial

neural networks. In 2013 IEEE 25th International Conference on Tools with Artificial Intelligence (pp. 159-166). IEEE.

9. Kadar, M., Sarraipa, J., Guevara, J. C., & Restrepo, E. G. Y. (2018, June). An Integrated Approach for Fighting Dropout and Enhancing Students' Satisfaction in Higher Education. In Proceedings of the 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion (pp. 240-247).
10. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.
11. Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory (pp. 144-152).
12. Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.
13. Paul, A., Mukherjee, D. P., Das, P., Gangopadhyay, A., Chintha, A. R., & Kundu, S. (2018). Improved random forest for classification. IEEE Transactions on Image Processing, 27(8), 4012-4024.
14. Dong, G., & Liu, H. (Eds.). (2018). Feature engineering for machine learning and data analytics. CRC Press.
15. Kim, Y., & Kim, Y. (2022). Explainable heat-related mortality with random forest and SHapley Additive exPlanations (SHAP) models. Sustainable Cities and Society, 79, 103677.
16. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
17. Liu, J., Danait, N., Hu, S., & Sengupta, S. (2013, December). A leave-one-feature-out wrapper method for feature selection in data classification. In 2013 6th International Conference on Biomedical Engineering and Informatics (pp. 656-660). IEEE.
18. Abid, A., Abdalla, A., Abid, A., Khan, D., Alfozan, A., & Zou, J. (2019). Gradio: Hassle-free sharing and testing of ml models in the wild. arXiv preprint arXiv:1906.02569.
19. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.
20. Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: data mining and knowledge discovery, 9(3), e1301.