Journal of Electrical Systems and Information Technology

RESEARCH Open Access

Diabetes type 2 classification using machine learning algorithms with up-sampling technique

Mariwan Ahmed Hama Saeed*

*Correspondence: mariwan.hamasaeed@uoh.edu.iq; mariwan.ahmedh@gmail.com

College of Basic Education, University of Halabja, Halabja, Iraq

Abstract

Recently, the rate of chronic diabetes disease has increased extensively. Diabetes increases blood sugar and other problems like blurred vision, kidney failure, nerve problems, and stroke. Researchers for predicting diabetes have constructed various models. In this paper, gradient boosting classifier, AdaBoost classifier, decision tree classifier, and extra trees classifier machine learning models have been utilized for identifying chronic diabetes disease. The models analyze the PIMA Indian Diabetes dataset (PIMA) and Behavioral Risk Factor Surveillance System (BRFSS) diabetes datasets to classify patients with positive or negative diagnoses. 80% of the datasets are used as training data and 20% as testing data. The extra trees classifier with an area under curve of 0.96% for PIMA and 0.99% for BRFSS datasets outperformed other models. Therefore, it is suggested that healthcare providers can use the ETC model to predict chronic disease.

Keywords: Diabetes, Diabetes type 2, Machine learning, Extra tree classifier, Up-sampling

Introduction

Diabetes is a widespread disease that happens in patients without enough insulin hormone. Human blood sugar is controlled by insulin [1, 2]. Increased blood sugar over time without control leads the body to serious health problems like lower limb amputation, blindness, and heart attacks [1–3]. In 2019, [3] estimated 1.9 million deaths because of diabetes, and it is the leading cause of death worldwide. In Early diagnosis, doctors analyze diabetes by using their information, but sometimes it might be inaccurate. Healthcare providers collect large amounts of data that cannot be used for effective decisions about diabetes disease [4]. Therefore, predicting and measuring the risk of diabetes disease using computer-based models can crucially reduce healthcare costs [5].

Numerous kinds of research have been devoted to modeling different diseases, including diabetes. Most of them trained the models using various features, for example, pregnancies, gender, age, and BMI [6–8].

Lu et al. [5] utilized support vector machine, logistic regression, K-nearest neighbors, Naïve Bayes, decision tree, random forest, XGBoost machine learning, and artificial

neural network deep learning models for predicting diabetes. They stated that RF was the best model, with an accuracy of 91, for predicting diabetes type 2. Various machine learning techniques were evaluated by [6] for classifying diabetes using PIMA diabetes dataset. Linear discriminant analysis was selected by [6] with an accuracy of 77 as the best model versus the other used machine learning techniques. Artificial neural networks, ontology classifiers, K-nearest neighbors, support vector machine, Naive Bayes, decision tree, and logistic regression were utilized to classify diabetes [9]. The ontology classifier with an accuracy of 77.5 was nominated as the best classification model. Farajollahi et al. [10] examined performance comparisons of XGBoost, decision tree, random forest, AdaBoost support vector machine, and logistic regression for diabetes diagnosis. They explained that AdaBoost has the most accuracy of 83 among other models. SVM and ANN are used by [11] for predicting the diagnosis of diabetes. Their model's accuracy was 94.87, higher than the other published works. RF and SVM algorithms were compared by [12] for diabetes prediction using feature selection and dimensionality reduction. Their work's accuracy was 81.4 and 83 for the used models. Other diabetes datasets have been utilized by researchers for classifying diabetes such as the Behavioral Risk Factor Surveillance System (BRFSS) [13] dataset which is considered by [14–16]. Furthermore, the above research utilized most machine and deep learning models with different accuracies in diabetes disease prediction.

The primary intent of this paper is to select the best machine learning model among four different models, which are not or less used in the literature for diabetes disease classification. PIMA and BRFSS datasets are studied in this paper using DTC, AdaBoost, GBC, ETC machine learning classifiers. The evaluation of the used classifiers is well organized. As a result of the work, the extra trees classifier provides superior

ROC of 0.96% for PIMA and 0.99% for BRFSS datasets versus the other used algorithms in diabetes classification. The rest of this work is used for: Materials and methods are explained in part two. Results and discussion are presented in part three. Part four summarizes significant conclusions.

Materials and methods

The proposed framework is illustrated in Fig. 1 and is divided into five phases. Jupiter Notebook with Python is used for the entire implementation of the model. The PIMA dataset has been analyzed using Sklearn, Matplotlib, Pandas, and Numpy packages.

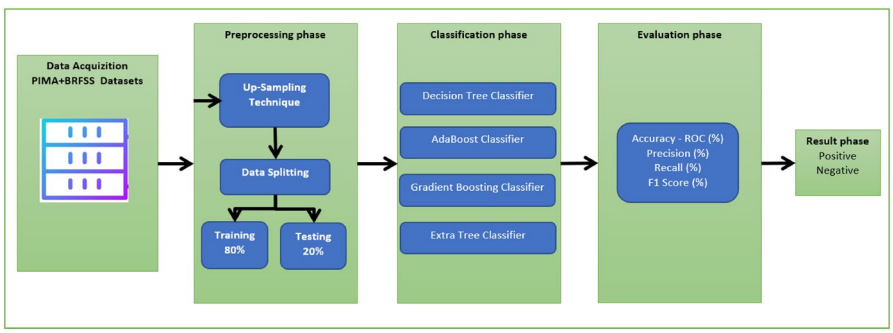

Fig. 1 The proposed framework for diabetes classification using machine learning models

Dataset

PIMA dataset

PIMA is a popular dataset in diabetes disease classification. It is from the NIDDK Institute [17]. It includes medical predictors of blood pressure, pregnancies, BMI, skin thickness, diabetes pedigree function, insulin, age, glucose, and outcome label. The label predicts positive and negative patients with a diagnosis of diabetes. The dataset initially includes imbalanced 769 samples, and 15 rows of the PIMA are presented in Table 1.

BRFSS dataset

Diabetes dataset from the Behavioral Risk Factor Surveillance System (BRFSS) gathered by the Centers for Disease Control and Prevention (CDC) [13]. It includes 253,680 samples with 21 feature variables such as a smoker, stroke, heart disease or attack, physical activity, fruits, sex, education, and income. The 15 rows of the dataset are presented in Table 2.

Preprocessing

The datasets should be preprocessed before applying them to classifiers. The outliers are removed from the datasets. The outcome and diabetes labels of PIMA and BRFSS datasets are not balanced. Unbalancing data decrease the accuracy of the classifiers. To mitigate this, the up-sampling technique [18] has been used to balance both datasets. After that, 80% of the datasets are used as training data and 20% as testing data randomly using the train-test-split function.

Table 1 15 Rows of the PIMA Dataset Sample

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1

Table 2 15 Rows of BRFSS dataset sample

High BP	High Chol	CholCheck	BMI	Smoker	Stroke	Heart Disease or Attack	Phys Activity	Fruits	Veggies	HvyAlcohol Consump	Any Health care	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income	Diabetes
1	1	1	40	1	0	0	0	0	1	0	1	0	5	18	15	1	0	9	4	3	0
0	0	0	25	1	0	0	1	0	0	0	0	1	3	0	0	0	0	7	6	1	0
1	1	1	28	0	0	0	0	1	0	0	1	1	5	30	30	1	0	9	4	8	0
1	0	1	27	0	0	0	1	1	1	0	1	0	2	0	0	0	0	11	3	6	0
1	1	1	24	0	0	0	1	1	1	0	1	0	2	3	0	0	0	11	5	4	0
1	1	1	25	1	0	0	1	1	1	0	1	0	2	0	2	0	1	10	6	8	0
1	0	1	30	1	0	0	0	0	0	0	1	0	3	0	14	0	0	9	6	7	0
1	1	1	25	1	0	0	1	0	1	0	1	0	3	0	0	1	0	11	4	4	0
1	1	1	30	1	0	1	0	1	1	0	1	0	5	30	30	1	0	9	5	1	1
0	0	1	24	0	0	0	0	0	1	0	1	0	2	0	0	0	1	8	4	3	0
0	0	1	25	1	0	0	1	1	1	0	1	0	3	0	0	0	1	13	6	8	1
1	1	1	34	1	0	0	0	1	1	0	1	0	3	0	30	1	0	10	5	1	0
0	0	1	26	1	0	0	0	0	1	0	1	0	3	0	15	0	0	7	5	7	0
1	1	1	28	0	0	0	0	0	1	0	1	0	4	0	0	1	0	11	4	6	1
0	1	1	33	1	1	0	1	0	1	0	1	1	4	30	28	0	0	4	6	2	0
1	0	1	33	0	0	0	1	0	0	0	1	0	2	5	0	0	0	6	6	8	0

Machine learning classification

In this work, gradient boosting classifier (GBC), decision tree classifier (DTC), extra trees classifier (ETC), and AdaBoost classifier (ABC) machine learning algorithms for classification problems are examined.

These models have been selected because they have recorded the highest accuracies and need less computing power than other machine and deep learning models. The extra trees classifier is chosen because it well predicted diabetes disease with area under curve accuracy of 96% for PIMA and 99% for the BRFSS compared to the DTC, GBC, and ABC.

Decision tree classifier (DTC)

It can perform regression, classification, and multioutput tasks because it is a powerful and versatile machine learning algorithm. It is also called the primary ensemble learning model and can fit a complex and large amount of data. It uses trees and validates values from the root till the last node [5, 19, 20]. This algorithm is used in this paper, and it recorded a ROC accuracy of 0.78 for the PIMA and 0.92 for the BRFSS datasets.

AdaBoost classifier (ABC)

It is a popular learning algorithm used in machine learning performance enhancement. Its base classifier is trained by the initial weight of $Xi = 1/n$ [21]. The classifier's performance depends on the former classifier. n means of training number instances, and Xi denotes the training sample. The final classifier is produced after the training of the base classifiers [21]. This model recorded a ROC accuracy of 0.83 for the PIMA and 0.82 for the BRFSS datasets.

Gradient boosting classifier (GBC)

Gradient boosting classifier like AdaBoost is one of the learning algorithms. Its predecessor was corrected after adding predictors sequentially to the ensemble[21]. It fits the new predictor's residual errors by the new predictor. It can be used in many areas such as ecology and Web search ranking [21–23]. After the extra trees classifier, this model recorded an accuracy ROC of 0.90 for the PIMA and 0.82 for the BRFSS datasets.

Extra trees classifier (ETC)

Extra trees classifier is the bagging machine learning algorithm where training dataset samples implement random trees [20, 22]. In machine learning, extra trees classifier and extra tree regressor are responsible for constructing extra trees, mitigating overfitting, and improving the classification accuracy [20, 24]. Lastly, this model is proposed by this paper since it significantly predicted diabetes disease with area under curve accuracy of 96% and 99% for both datasets compared to the published works in the literature.

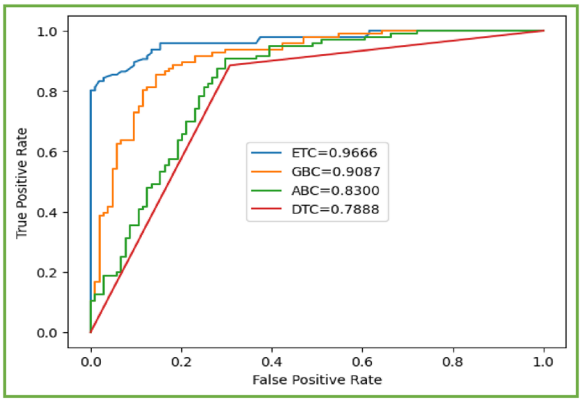

Fig. 2 The ROC plot of all classifiers for the PIMA dataset

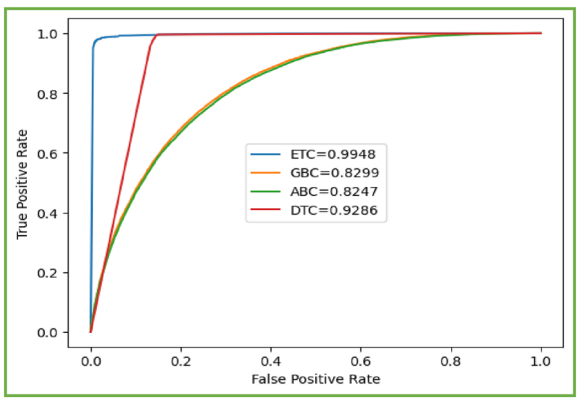

Fig. 3 The ROC plot of all classifiers for the BRFSS dataset

Evaluation metric

The models are evaluated using different evaluation metrics, including receiving operating characteristic (ROC) curve, accuracy, F1-Score, precision, and recall [20, 23] which are presented in Figs. 2, 3, and Table 3, respectively.

Accuracy is the number of observations predicted correctly

Accuracy = (True Positive + True Negative)/(True Positive + True Negative + False Positive + False Negative).

Precision is the number of each observation predicted to be positive that is positive.

Precision = True Positive/(True Positive + False Positive).

The recall is the number of each positive observation that is genuinely positive.

Recall = True Positive/(True Positive + False Negative).

where

True positive is correctly predicted by the classifier as a positive class

True negative is correctly predicted by the classifier as a negative class

False positive is incorrectly predicted by the classifier as a positive class

False negative is incorrectly predicted by the classifier as a negative class

Table 3 The classification report for the selected models and datasets

Classifiers	PIMA Dataset				BRFSS Dataset		
	Diabetes	Precision	Recall	f1-score	Precision	Recall	f1-score
Extra Tree Classifier	Negative	0.92	0.87	0.89	0.99	0.94	0.96
	Positive	0.86	0.92	0.89	0.94	0.99	0.97
	Accuracy 0.89				Accuracy 0.96		
	Unweighted Avg	0.89	0.89	0.89	0.97	0.96	0.96
	Weighted avg	0.89	0.89	0.89	0.97	0.96	0.96
Decision Tree Classifier	Negative	0.90	0.71	0.80	0.99	0.86	0.92
	Positive	0.75	0.92	0.82	0.87	0.99	0.93
	Accuracy 0.81				Accuracy 0.92		
	Unweighted Avg	0.82	0.81	0.81	0.93	0.92	0.92
	Weighted avg	0.83	0.81	0.81	0.93	0.92	0.92
AdaBoost Classifier	Negative	0.79	0.76	0.77	0.75	0.73	0.74
	Positive	0.75	0.78	0.77	0.74	0.77	0.75
	Accuracy 0.77				Accuracy 0.75		
	Unweighted Avg	0.77	0.77	0.77	0.75	0.75	0.75
	Weighted avg	0.77	0.77	0.77	0.75	0.75	0.75
Gradient Boosting Classifier	Negative	0.88	0.81	0.84	0.77	0.71	0.74
	Positive	0.81	0.89	0.85	0.73	0.79	0.76
	Accuracy 0.84				Accuracy 0.75		
	Unweighted Avg	0.85	0.85	0.84	0.75	0.75	0.75
	Weighted avg	0.85	0.84	0.84	0.75	0.75	0.75

$F1 =$ harmonic mean of precision and recall $= 2 \times [(\text{Recall} \times \text{Precision})/(\text{Recall} + \text{Precision})]$.

The results of precision, accuracy, recall and F1-Score for the classifiers are shown in Table 3.

The Receiving Operating Characteristic (ROC) curve compares false and true positives at each threshold. Its plot for the actual false positive versus the positive rates of all classifiers is shown in Figs. 2 and 3.

Higher values for the mentioned five metrics demonstrate better performance for the classifier.

Results and discussion

Figures 2 and 3 show the ROC results of the four machine learning models using testing/validation data for the datasets. Extra trees classifier has the highest ROC of 0.96% for PIMA and 0.99% for BRFSS datasets respectively, whereas the decision tree and AdaBoost classifiers have the lowest ROC of 0.7888 and 0.8247 for the PIMA and the BRFSS datasets, respectively.

Table 3 shows the classification report for the chosen models and datasets, including accuracy, recall, precision, F1-Score, weighted, and unweighted accuracy. It is

Table 4 Comparison with the state of the art for PIMA and BRFSS datasets

No.	Authors	Models	PIMA—Accuracy	Authors	Models	BRFSS—Accuracy
1	Lu et al. [5]	Random Forest	84.95	Dinh et al. [16]	eXtreme Gradient Boost	95
2	Mujumdar et al. [6]	Linear Discriminate Analysis	77	Maniruzzaman et al. [15]	Linear Regression + RandomForest	94.25%
3	Massari et al. [9]	Ontology classifiers and SVM	77.5	Nadeem et al. [14]	SVM + ANN	94.67%
4	Farajollahi et al. [10]	AdaBoost Classifier	83			
5	Sivaranjani et al. [12]	Random Forest	83			
	Proposed Model	Extra Tree Classifier	89	Proposed Model	Extra Tree Classifier	96

worth noting that the BRFSS dataset is better recognized than the PIMA dataset by the extra trees classifier.

For the PIMA dataset, positive and negative classes with the highest recognition rate of 92% and 87% were detected by the extra tree classifier, whereas the AdaBoost and decision tree classifiers identified the positive and negative classes with the lowest recognition rate of 78 and 71%, respectively.

For the BRFSS dataset, both extra tree and decision tree classifiers have the greatest recognition rate of 99% for the positive class and 94% for the negative class, correspondingly. However, the gradient boosting and AdaBoost classifiers had the lowest recognition rates, with 71% for the negative class and 77% for the positive class.

Table 4 compares the four utilized models with the state-of-the-art research for predicting diabetes disease using PIMA and BRFSS datasets. The largest accuracy of 84.95% was obtained by Lu et al. for the PIMA and an accuracy of 95% was gained by Dinh et al. for the BRFSS datasets. Nevertheless, the proposed model has an excellent prediction with an accuracy of 89% for the PIMA and 96% for the BRFSS datasets among the mentioned papers in the literature.

Finally, I observed that using the up-sampling strategy to balance the previously described imbalanced datasets with the extra tree classifier for diabetes detection yielded the greatest recognition rates among published works in the literature.

Conclusion

Machine learning techniques are considered crucial for disease prediction. In this paper, four machine learning models have been proposed for the classification of diabetes type 2. The PIMA and BRFSS datasets have been used with the help of the up-sampling technique for balancing the dataset. The extra trees classifier with an area under curve of 0.96% for PIMA and 0.99% for BRFSS outperformed other models. The findings of this research confirm that healthcare providers can use the ETC model for predicting chronic diseases. Deep learning models can be utilized in future work to predict other diseases. Also, using data fusion and hybrid models will be studied as well.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s43067-023-00074-5.

Additional file 1: Diabetes binary health indicators BRFSS2015 dataset.

Additional file 2: PIMA INDIAN diabetes dataset.

Acknowledgements
My appreciation goes to my wife who is always helpful.

Author contributions
Mariwan Ahmed Hama Saeed has written the whole paper.

Funding
Not received.

Availability of data and materials
Available on Request.

Declarations

Competing interests
The author declares that there are no competing interests.

Received: 7 October 2022 Accepted: 10 January 2023
Published: 8 February 2023

References

1. Centers for Disease Control and Prevention, "What is diabetes? | CDC." https://www.cdc.gov/diabetes/basics/diabetes.html (accessed Aug. 28, 2022)
2. Mayo Clinic Staff (2022) Diabetes - Symptoms and causes - Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444 (accessed Aug. 28, 2022)
3. World Health Organization (2022) Diabetes." https://www.who.int/news-room/fact-sheets/detail/diabetes (accessed Aug. 28, 2022).
4. Naz H, Ahuja S (2020) Deep learning approach for diabetes prediction using PIMA Indian dataset. J Diabetes Metab Disord 19(1):391–403. https://doi.org/10.1007/S40200-020-00520-5
5. Lu H, Uddin S, Hajati F, Moni MA, Khushi M (2022) A patient network-based machine learning model for disease prediction: the case of type 2 diabetes mellitus. Appl Intell. https://doi.org/10.1007/s10489-021-02533-w
6. Mujumdar A, Vaidehi V (2019) Diabetes prediction using machine learning algorithms. Proc Comput Sci. https://doi.org/10.1016/j.procs.2020.01.047
7. Sahoo AK, Pradhan C, and Das H (2020) Performance evaluation of different machine learning methods and deep-learning based convolutional neural network for health decision making. In: Studies in Computational Intelligence, vol. SCI 871, https://doi.org/10.1007/978-3-030-33820-6_8
8. Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G (2020) Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Sci Rep. https://doi.org/10.1038/s41598-020-68771-z
9. el Massari H, Mhammedi S, Sabouri Z, and Gherabi N (2022) Ontology-based machine learning to predict diabetes patients. In: Lecture notes in networks and systems, vol. 357 LNNS. https://doi.org/10.1007/978-3-030-91738-8_40
10. Farajollahi B, Mehmannavaz M, Mehrjoo H, Moghbeli F, Sayadi MJ (2021) Diabetes diagnosis using machine learning. Front Health Inform. https://doi.org/10.30699/fhi.v10i1.267
11. Ahmed U et al (2022) Prediction of diabetes empowered with fused machine learning. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3142097
12. Sivaranjani S, Ananya S, Aravinth J, and Karthika R (2021) Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction. In: 2021 7th international conference on advanced computing and communication systems, ICACCS 2021. https://doi.org/10.1109/ICACCS51430.2021.9441935
13. Diabetes Health Indicators Dataset | Kaggle. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?resource=download (accessed Nov. 26, 2022)
14. Nadeem MW, Goh HG, Ponnusamy V, Andonovic I, Khan MA, Hussain M (2021) A fusion-based machine learning approach for the prediction of the onset of diabetes. Healthcare 9(10):1393. https://doi.org/10.3390/HEALTHCARE9101393
15. Maniruzzaman M, Rahman MJ, Ahammed B, Abedin MM (2020) Classification and prediction of diabetes disease using machine learning paradigm. Health Inf Sci Syst 8(1):1–14. https://doi.org/10.1007/S13755-019-0095-Z/TABLES/13
16. Dinh A, Miertschin S, Young A, Mohanty SD (2019) A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak 19(1):211. https://doi.org/10.1186/s12911-019-0918-5
17. National Institute of Diabetes and Digestive and Kidney Diseases (2022) Pima Indians Diabetes - dataset by uci | data.world. https://data.world/uci/pima-indians-diabetes (accessed Aug. 28, 2022)

18. Brownlee J (2020) Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning. Machine Learning Mastery. https://books.google.pt/books?id=jaXJDwAAQBAJ
19. Jiang H (2021) Machine learning fundamentals : a concise introduction. https://books.google.iq/books?id=RzVfzgEACAAJ
20. Géron A (2019) Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems. https://books.google.iq/books?id=HHetDwAAQBAJ
21. Rafatirad S, Homayoun H, Chen Z, and Pudukotai Dinakarrao SM (2022) Machine learning for computer scientists and data analysts. https://doi.org/10.1007/978-3-030-96756-7
22. Brownlee J (2017) Machine learning mastery with python: understand your data, create accurate models and work projects end-to-end, Machine Learning Mastery, vol. 91
23. Albon C (2018) Machine learning with Python cookbook : practical solutions from preprocessing to deep learning. https://books.google.iq/books?id=VucltAEACAAJ
24. Scikit-learn (2022) sklearn.ensemble.ExtraTreesClassifier — scikit-learn 1.1.2 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html (accessed Aug. 29, 2022)

Publisher's Note