E-ISSN 2586-9027
Homepage : https://tci-thaijo.org/index.php/SciTechAsia Science & Technology Asia
Vol.30 No.3 July - September 2025 Page: [51-59]
Original research article

Cup-lump Rubber Quality Estimation from Historical Weather Data Using Machine Learning Regressors

Charisa Areewattana[1], Nattapon Pramotkul[2], Tanakorn Chanjatunat[2], Somrudee Deepaisarn[1, 3, *]

[1] *Sirindhorn International Institute of Technology, Thammasat University Pathum Thani 12120, Thailand*
[2] *Southland Rubber Company Ltd., Sonkhla 90110, Thailand*
[3] *Research Unit in Sustainable Electrochemical Intelligent, Thammasat University, Pathum Thani 12120, Thailand*

Received 14 June 2025; Received in revised form 25 July 2025
Accepted 19 August 2025; Available online 30 September 2025

ABSTRACT

Rubber is a vital raw material for many industries, with rubber trees requiring specific climate conditions to thrive. Thailand is a major producer, particularly of cup-lump rubber, widely used in automobile tires. Rubber quality is assessed by the dry rubber content (DRC) percentage, which represents the usable portion after processing. Traditional methods of measuring DRC often depend on human judgment, creating potential bias. This study introduces a machine-learning approach to predict DRC percentages using historical weather data. Building on earlier research that employed statistical models, this work incorporates a broader range of weather-based features and modern regression algorithms. Weather variables—including temperature, precipitation, wind speed, and sunlight duration—were obtained through the Open-Meteo API and averaged over rolling time windows spanning 3 to 50 days. These features were combined with 2,034 in-house DRC records supplied by Southland Rubber Company Ltd., resulting in a dataset of 124 features. Multiple regression models were evaluated with Scikit-learn, including XGBoost, LightGBM, and Random Forest. Among them, XGBoost delivered the best results, achieving an R^2 score of 0.7450 and a root mean square error of 3.23%. The findings show that machine learning can reliably predict rubber quality from weather trends. This provides manufacturers with an objective, data-driven tool to improve production planning, resource allocation, and quality control.

Keywords: Dry Rubber Content (DRC); Machine learning; Regression; Rubber quality; Weather data

***Corresponding author:** somrudee@siit.tu.ac.th **doi:** 10.14456/scitechasia.2025.46

1. Introduction

The rubber tree (*Hevea brasiliensis*) is challenging to grow due to its stringent environmental requirements. It thrives only in areas with consistently warm and humid climates, stable temperatures, high annual rainfall, and fertile, well-drained soils. Even small changes, such as droughts, floods, or temperature swings, can significantly impact the tree's health and latex quality. Due to specific requirements, rubber cultivation is supported in some regions of the world, primarily in the equatorial belt, including Southeast Asia, parts of South America, and specific regions in Africa. This special geography permits strategic rubber production for the countries involved. Rubber is not only a natural resource—it is a foundational material that supports the functioning of countless industries. The unique properties of rubber, including elasticity, waterproofing, heat resistance, and durability, make it essential for a wide range of applications. For example, in the automotive industry, rubber is used in tires, hoses, and suspension components. In the medical field, gloves, tubing, and various other protective equipment are used—everyday objects such as phone cases, elastic bands, and floor mats. Beyond visible uses, rubber often works behind the scenes in industrial machines, construction materials, and electronics. Rubber plays a critical role in global supply chains and supports the foundation of modern life. Without rubber, many of the conveniences and systems we rely on daily would be difficult—if not impossible—to maintain.

Dry rubber content (DRC) percentage in cup-lump rubber is a critical indicator of rubber quality. It refers to the percentage of solid rubber present in the latex or solidified rubber mass after removing water and other contaminants. A higher percent DRC indicates a better quality and higher economic value of the raw rubber material. Traditionally, determining the DRC of cup-lump rubber has relied heavily on human expert labor with manual techniques, which can be labor-intensive, time-consuming, and prone to human error, particularly when applied in field conditions without advanced equipment.

Thailand holds a major position as a global producer and exporter of natural rubber. However, this production is sensitive to fluctuations in weather conditions. Several observed studies [1-5] have consistently demonstrated that meteorological variables—especially rainfall, temperature, and relative humidity—have a profound effect on both the physiological functions and latex yield of *Hevea brasiliensis*.

Several studies have examined the relationship between environmental variables and rubber productivity. Zhang et al. (2019) applied multiple regression analysis using tree-related factors—such as clone, age, tapping method, and location—to predict the DRC, demonstrating the utility of statistical methods in rubber agriculture [5]. While effective, such models may lack flexibility when dealing with large-scale or nonlinear data structures.

Recent advancements have introduced machine learning (ML) into this domain. Wadugoda et al. (2025), for instance, combined meteorological parameters with satellite-derived indices to estimate rubber yield in Sri Lanka, demonstrating how ML can adapt to complex, region-specific agricultural data [6]. Similarly, Puttipipatkajorn (2021) implemented a spectroscopic approach integrated with machine learning algorithms to evaluate DRC in cup-lump rubber, emphasizing the potential of sensor-based precision analytics [7].

Compared to prior studies that rely on spectral or biological data, our weather-only approach yields slightly lower predictive performance (R^2 = 0.7450) but offers greater accessibility and scalability. By eliminating the need for specialized equipment, it becomes practical for large plantations and resource-limited settings. While weather-based models are less precise, they can serve as a cost-effective, non-invasive alternative or an early-stage screening tool. However, a key limitation is geographic generalizability. Since the DRC data in this study were collected from plantations in and around Bueng Kan province, model performance may vary when applied to regions with different climatic or soil conditions.

This research employs a novel approach, focusing on the use of machine learning to predict DRC based solely on weather data obtained from the OpenMeteo API [8], averaged over multiple temporal windows (3-50 days). Data for DRC percentages and time-specific weather-related variables are recorded internally at Southland Rubber Company Ltd., Thailand. The final dataset comprises 2,034 observations, and 124 input features were used to build regression models. The XGBoost regressor was selected as a suitable candidate for predicting DRC percentages. It highlights the practical potential of machine learning in leveraging weather and production data to enhance rubber quality estimation. Models can be utilized to predict and suggest weather conditions that impact production efficiency, enhance quality control, and inform strategic decisions regarding labor and resource management for the rubber industry.

2. Methodology

This section outlines the dataset acquisition and machine learning regression experiment to select the appropriate model. Fig. 1 illustrates the overall workflow for the training and testing phases of cup-lump dry rubber content (DRC) percentage estimation, utilizing historical weather data as input for the machine learning regression models.

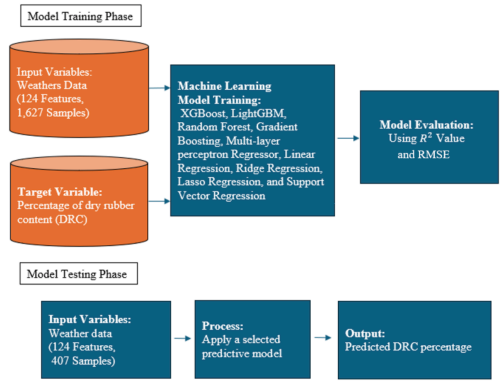

Fig. 1. Workflow diagram.

2.1 Dataset

Data span from December 3, 2018, to January 20, 2025. The dataset consists of 2,034 observations and 17 original weather-related features extracted from the Open-Meteo API [8] as listed in Table 1 (1-17), averaged over multiple historical temporal windows of 3, 5, 7, 10, 15, 30, and 50 days, incorporating 119 derived features and 5 fixed feature data (latitude, longitude, Day, Month, Year) as listed in Table 1 (18-22)

All these features are input variables for the regression model. In addition, date information (i.e., date, month, and year) and the geographic data (latitude and longitude of the rubber tree garden), for each observation from in-house records are included as input variables. This makes up 124 input-feature set.

DRC percentages as listed in Table 1 (feature No. 23) sourced directly from the internal records of Southland Rubber

Table 1. Dataset description compounding features name, description, and basic statistic.

No.	Name	Description	Unit	Mean	Median	Mode	Max	Min	SD
1	weather_code	The most severe weather condition on a given day		32.0611	32.25	34.4929	65	0	19.541
2	temperature_ 2m_max	Maximum daily air temperature at 2 meters above ground	°C	30.405	30.284	30.2843	40.45	21.475	2.195
3	temperature_ 2m_min	Minimum daily air temperature at 2 meters above ground	°C	21.4847	21.637	21.637	29.725	10.725	2.749
4	temperature_ 2m_mean	Mean daily air temperature at 2 meters above ground	°C	25.5921	25.58	25.58	34.8666	15.55	2.264
5	apparent_ temperature_max	Maximum daily apparent temperature	°C	34.0661	34.079	34.7907	44.2333	20.075	3.539
6	apparent_ temperature_min	Minimum daily apparent temperature	°C	23.8888	2408689	24.1643	34.2419	7.975	4.327
7	apparent_ temperature_mean	Mean daily apparent temperature	°C	28.4674	28.5935	28.5935	37.5833	13.1	3.729
8	precipitation_sum	Sum of daily precipitation	mm	4.46494	4.2854	0	49.725	0	5.065
9	rain_sum	Sum of daily rain	mm	4.46494	4.2854	0	49.725	0	5.065
10	precipitation_hours	The number of hours with rain	hr.	4.55776	4.4882	0	23	0	4.333
11	daylight_duration	Number of seconds of daylight per day	sec.	42894.1	42875.8	42934.2	47604.9	39707.5	1987.31
12	sunshine_duration	Daily sunshine duration is calculated based on direct normalized irradiance exceeding 120 W/m^2	sec.	33949.4	33816.3	33147.2	43292.9	5054.1	4480.87
13	wind_speed_ 10m_max	Maximum wind speed on a day	km/h	13.8689	13.8327	13.8327	29.875	5.75	2.657
14	wind_gusts_ 10m_max	Maximum wind gusts on a day	km/h	30.4784	30.4526	30.4526	58.225	16.025	4.478
15	wind_direction_ 10m_dominant	Dominant wind direction	°	124.505	123.705	130.337	356.25	2.75	54.476
16	shortwave_ radiation_sum	The sum of solar radiation on a given day in Megajoules	MJ/m^2	18.2497	18.168	17.8643	25.5	6.82	2.048
17	et0_fao_ evapotranspiration	Daily sum of ET_0 Reference Evapotranspiration of a well watered grass field	mm	4.17227	4.10875	4.06127	8.5512	1.55545	0.6336
18	Latitude	Location	°	16.0974	16.323	16.32	20.04	6.665	2.297
19	Longitude	Location	°	102.701	103.166	102.476	104.918	99.344	1.313
20	Day	Temporal Data		15.9247	16	7	31	1	8.593
21	Month	Temporal Data		6.6258	7	1	12	1	3.911
22	Year	Temporal Data		2021.97	2022	2022	2025	2018	1.321
23	TB_Receive_ RM_PercentDRC	DRC percentage ground truth	%	64.01	63	61	84	50	6.265

Company Ltd., Thailand is the target variable for the regression models to predict. Table 1 (feature No. 1–17) presents the original weather-based variables, derived using statistical measures—mean, median, mode, maximum, minimum, and standard deviation—calculated across average multiple time windows, as discussed in the previous paragraph.

2.2 Experimental set-up for machine learning regressors

The dataset contains 2,034 observations, with 80% (1,627) used for training and 20% (407) for testing. A fixed random seed (80) was set to ensure reproducibility. Given the dataset's limited size, no separate validation set was created; instead, internal cross-validation within the training set was used for hyperparameter tuning and model

selection to prevent overfitting and maintain evaluation fairness.

Fixed seed ensures that random operations—such as data shuffling and splitting—yield the same results each time the code is executed, which is essential for reliable model evaluation and fair comparison across experiments.

A variety of machine learning regression models were built and evaluated to predict dry rubber content (DRC). Models include Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Random Forest, Gradient Boosting Regressor, Multi-layer Perceptron (MLP), Linear Regression, Ridge Regression, Lasso Regression, and Support Vector Regression (SVR).

R^2 score and root mean squared error (RMSE) were used as model evaluation metrics. R^2 can assess the agreement between the predictions and ground truth values, and RMSE measures the goodness of fit of the model prediction.

3. Results and Discussion

Experiments, as described in Section 2.2, were conducted to inform model selection and parameter optimization. The performances of a variety of 9 machine learning regression models in predicting dry rubber content (DRC) were compared under the same standardized preprocessing pipeline to ensure consistency and fairness in the comparative analysis. R^2 test and RMSE of test set, quantifies the proportion of variance in the target variable (i.e., DRC) explained by the prediction as a result of input variables within the model, and the root mean squared error (RMSE), which measures the average deviation of the predicted values from the actual values. Among all models tested, XGBoost achieved the highest R^2 test score, and also recorded one of the lowest RMSE values. And the second is LightGBM lean to adjust those 2 models.

LightGBM the final tuned configuration is as follows: objective = 'reg:squarederror', learning_rate = 0.05, max_depth = 8, subsample = 0.9, colsample_bytree = 0.9, n_estimators = 1600, gamma = 0, reg_alpha = 0.1, and reg_lambda = 1.0. Show in Table 2 achieving an R^2 score of 0.7387 on the testing set, with a mean squared error (MSE) of 10.75 and root mean squared error (RMSE) of 3.28; and an R^2 score of 0.9840 on the training set, with an MSE of 0.62 and RMSE of 0.79.

The XGBoost the final tuned configuration is as follows: objective = 'reg:squarederror', learning_rate = 0.05, max_depth = 8, subsample = 0.9, colsample_bytree = 0.9, n_estimators = 1600, gamma = 0, reg_alpha = 0.1, and reg_lambda = 1.0.

In Table 2 Xgboost model achieving an R^2 score of 0.7450 on the testing set, with a mean squared error (MSE) of 10.01 and root mean squared error (RMSE) of 3.16; and an R^2 score of 0.9867 on the training set, with an MSE of 0.52 and RMSE of 0.72.

Adjusting the parameters max_depth and n_estimators reveals a clear trend of overfitting, as illustrated in Fig. 2. exhibits a noticeable gap from the training performance—an indication of overfitting. Given that the dataset includes features related to weather conditions, the risk of overfitting is inherently high. This observation is consistent with the findings of Peisong Niu et al. (2025), who noted that "The problem of overfitting becomes more severe when dealing with limited and noisy real-world data, as models tend to memorize specific instances rather than generalize fundamental patterns" [12].

Table 2. The table shows models performance both train test of R^2 Score, RMSE, MSE, and MAE are sorted by R^2 test Score (highest to lowest).

	Model	Test R^2	Train R^2	Test RMSE	Test MSE	Test RMSE	Train MSE	Test MAE	Train MAE
1	XGBoost	0.745	0.9867	3.1641	10.0118	0.7192	0.5173	1.941	0.2467
2	LightGBM	0.7387	0.984	3.278	10.7456	0.787	0.6194	2.0957	0.3676
3	Random Forest	0.6947	0.9408	3.5433	12.5552	1.5149	2.2949	2.2886	0.9897
4	Gradient Boosting	0.5291	0.645	4.4007	19.3657	3.7089	13.7562	3.0992	2.6983
5	MLP Regressor	0.2979	0.8237	5.3733	28.8724	2.6137	6.8316	3.8339	1.9177
6	Linear Regression	0.2365	0.4018	5.6036	31.3998	4.8149	23.1833	4.1621	3.6185
7	Ridge Regression	0.2351	0.3964	5.6085	31.4553	4.8367	23.3932	4.1454	3.6356
8	Lasso Regression	0.2279	0.2842	5.6351	31.7546	5.267	27.7408	4.1483	3.8422
9	Support Vector Regression	0.1761	0.2671	5.8208	33.8814	5.3295	28.4037	3.7569	3.3547

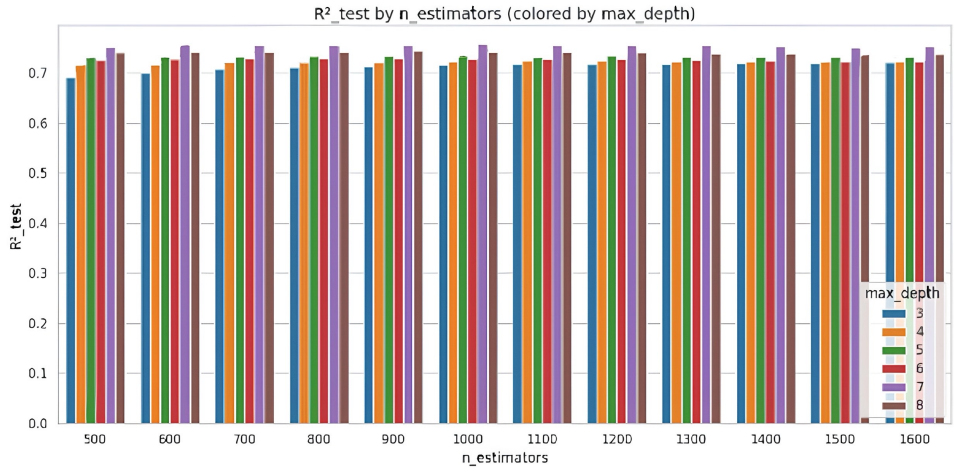

Fig. 2. Bar graph showing R^2 score in test set at varied n_estimate and max_depth parameeters.

An R^2 test set score of 0.7450 indicates that approximately 74.5% of the variance in the dry rubber content (DRC) percentage can be explained by the model, indicating the agreement between the predicted and the ground truth values. This suggests a strong correlation between the predicted values and the actual measurements. The relatively high R^2 value highlights the ability of the XGBoost algorithm to capture complex, nonlinear relationships within the dataset.

An analysis of the feature importance as seen in Fig. 3 reveals the top three most influential features: Longitude (lng) — the most important feature (0.0108) 30-day average precipitation hours — the number of hours with rainfall over the past 30 days 30-day average apparent temperature (mean) — the average "felt" temperature over the past 30 days ' Geographical location plays a significant role. The features lng and lat rank among the most important, suggesting that the physical location—possibly due to variations in local climate or soil type—strongly influences rubber yield.

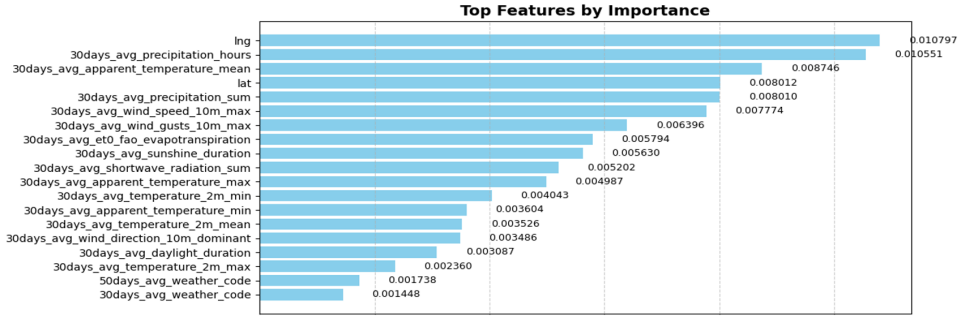

Fig. 3. The bar graph shows top 10 most important features contributing to the XGBoost model's predictions.

Recent 30-day data proves more relevant than longer timeframes, with features based on 30-day means consistently ranking higher than those based on 50-day periods. For instance, 30days_avg_weather_code outperforms 50days_avg_weather_code, suggesting that short-term weather conditions prior to measurement or harvest have a greater influence on DRC outcomes than longer-term trends.

Weather conditions—particularly rainfall and apparent temperature—are directly linked to latex yield. Rain-related variables such as 30days_avg_precipitation_hours and 30days_avg_precipitation_sum, along with apparent temperature indicators (mean, minimum, and maximum), rank among the top features. This implies that consistent rainfall and variations in perceived temperature substantially impact latex production in the region.

While few studies have directly examined the link between rubber quality and weather conditions in the preceding 30 days, both in-house experts at Southland Rubber Company Ltd. and prior research support this connection. Thaler et al. (2010) highlight the adverse effects of water scarcity and drought on rubber production [10], while Sinnarong et al. (2021) report a strong correlation between rainfall variability and yield [11]. These findings emphasize the combined importance of temporal (short- and long-term weather patterns) and spatial (geographic location) factors in predicting DRC percentages. Integrating these dimensions enables the model to capture complex interactions between environmental conditions and regional characteristics, ultimately enhancing rubber quality forecasting. This insight supports more informed decision-making for supply chain optimization, production planning, and quality control.

4. Conclusion

This study confirms the effectiveness of machine learning, particularly the XGBoost model, in predicting the dry rubber content (DRC) percentage in cup-lump rubber using historical weather data. The dataset was constructed from weather data obtained via the Open-Meteo API combined with in-house DRC percentage data from Southland Rubber Company Ltd., with a total of 2,034 observation records. Key features include temperature, precipitation, wind speed, and sunlight duration, averaged over various time windows ranging from 3 to 50 days to capture both short- and long-term weather pattern.

The XGBoost model outperformed other regression models, R^2 score of 0.7450 on the testing set, with a mean squared error (MSE) of 10.01 and root mean squared error (RMSE) of 3.16; and an R^2 score of 0.9867 on the training set, with an MSE of 0.52 and RMSE of 0.72. These results indicate a strong relationship between weather conditions and rubber quality. Feature importance analysis revealed that 30-day average weather variables and geospatial data (latitude and longitude) had the greatest impact on model performance.

In summary, this approach offers a consistent and data-driven alternative to subjective DRC percentage assessments. It enables rubber producers to enhance quality control, forecast product quality more reliably, and make better-informed production decisions based on weather trends.

Acknowledgements

The first author would like to express sincere gratitude to Southland Rubber Company Ltd. for providing the opportunity to undertake the internship and for providing valuable data and insights that greatly contributed to the success of this research, and Sirindhorn International Institute of Technology (SIIT) for the academic guidance, encouragement, and continuous support throughout this research project.

References

[1] Xayyaphet B, Kohkhoo R, Pakoktom T. The relationship between meteorological factors and yield of para rubber tree. In: Proc. 12th National Academic Conference, Kasetsart University, Kamphaeng Saen Campus, Thailand; 2020.

[2] Thamchoti P, Kongrit V. Adaptation of rubber production system to climate change in upper southern Thailand. J Agric Sci. 2021;15(3):45–58.

[3] Sangchanda N, et al. Effect of climate on growth and physiological characteristics of rubber tree in northeast Thailand. Thai J Plant Physiol. 2019;10(2):120–135.

[4] Thaiburi N, Sinnarong N, Autchariyapanitkul K, Nunthasen K. Impacts of climate change on rubber production in lower southern Thailand. Narathiwat Rajanagarindra University, College of Agriculture and Technology; year unknown.

[5] Evaluation of the impact of climatic factors on latex yield of Hevea brasiliensis. Int J Res Stud Agric Sci. 2017;3(5). doi:10.20431/2454-6224.0305004.

[6] Zhang F, Lin Y, Song Y, Zhang J. Prediction of dry rubber content in Hevea latex based on SPSS multiple regression model. IOP Conf Ser Mater Sci Eng. 2019;563:022030.

[7] Puttipipatkajorn A, Puttipipatkajorn A. Spectroscopic measurement approaches in evaluation of dry rubber content of cup lump rubber using machine learning techniques. Int J Agric Biol Eng. 2021;14(3):207–213.

[8] Wadugoda HWNO, Perera TANT, Abhiram G, Jayasinghe GY. Machine learning-based yield prediction for rubber (Hevea brasiliensis) cultivation. Uva Wellassa University and University of Ruhuna, Sri Lanka; 2025.

[9] Open-Meteo API. Available from: https://open-meteo.com

[10] Thaler P, Jeanneau C, Sotta B, Maillard M, Michaud A, Ramel J, Prioul J-P, Montpied T. Effects of drought and tapping for latex production on water relations of Hevea brasiliensis trees. Tree Physiol. 2008;28(12):1791–1799. doi:10.1093/treephys/28.12.1791.

[11] Thaiburi N, Sinnarong N, Autchariyapanitkul K, Nunthasen K. Impacts of climate change on rubber production in lower

southern Thailand. Narathiwat Rajanagarindra University, College of Agriculture and Technology; 2021.

[12] Niu P, Ma Z, Zhou T, Chen W, Shen L, Jin R, Sun L. Utilizing strategic pre-training to reduce overfitting: Baguan – a pre-trained weather forecasting model. arXiv Preprint. 2025;arXiv:2505.13873.

[13] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proc. 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. 2016. p. 785–794.

[14] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–2830.