https://dx.doi.org/10.4314/afsjg.v14i1.12

Prediction of Tea Production in Rwanda Using Data Mining Techniques

***C. Umutoni[1] and I. Ngaruye[2]**

[1] African Centre of Excellence in Data Science, University of Rwanda, Kigali, Rwanda
[2] Department of Mathematics, College of Science and Technology, University of Rwanda, Kigali, Rwanda

*Corresponding Author: tonirisse@gmail.com

Received: 23rd June 2021 Accepted: 20th January 2022

Abstract

Rwanda's main economic activity is agriculture, and tea is the country's most important cash crop. There has been extensive research on prediction of tea production in Rwanda but most of the methods applied were the traditional statistical analyzes with limited prediction capability. Data mining algorithm models, linear regression, K-Nearest Neighbor (KNN), Random Forest Regression, and Extremely Randomized Trees are discussed in this study to identify critical features in different domains to facilitate accurate prediction of tea production in Rwanda. In this study also, an identification of different factors which are strongly associated with tea production and developed data mining models for predicting tea production using training and test data from National Agricultural Export Development Board (NAEB) 2010-2019 is performed and PYTHON, R, and SPSS Version 25 softwares used in this study. The findings reveal that extra tree and random forest are the best model among the others to predict tea production in Rwanda.

Keywords: Tea production, Data mining, model accuracy, tea prediction

Prévision de la production de thé au Rwanda à l'aide de techniques d'exploration de données

Résumé

La principale activité économique du Rwanda est l'agriculture, et le thé est la culture de rente la plus importante du pays. De nombreuses recherches ont été menées sur la prédiction de la production de thé au Rwanda, mais la plupart des méthodes appliquées étaient des analyses statistiques traditionnelles avec une capacité de prédiction limitée. Les modèles d'algorithmes d'exploration de données, la régression linéaire, le K-Nearest Neighbor (KNN), la régression Random Forest et les arbres extrêmement randomisés sont discutés dans cette étude pour identifier les caractéristiques critiques dans différents domaines afin de faciliter la prédiction précise de la production de thé au Rwanda. Dans cette étude également, une identification des différents facteurs qui sont fortement associés à la production de thé et des modèles d'exploration de données développés pour prédire la production de thé en utilisant des données d'entraînement et de test du National Agricultural Export Development Board (NAEB) 2010-2019 est effectuée et les logiciels PYTHON, R, et SPSS Version 25 sont utilisés dans cette étude. Les résultats révèlent que l'arbre supplémentaire et la forêt aléatoire sont les meilleurs modèles parmi les autres pour prédire la production de thé au Rwanda.

Mots-clés: Production de thé, exploration de données, précision du modèle, prédiction du thé.

Introduction

Agriculture is Rwanda's main economic activity, with 70 percent of the people employed in this industry, whereas agriculture employs around 72 percent of Rwanda's working population. In general, Rwanda's GDP has grown at an average rate of 7% from 2014 to 2019, with a 0.6% decrease in 2020 due to the Covid -19 pandemic (NISR). Rwanda's agricultural sector accounts for 33% of the national GDP, with tea and coffee being major export crops contributing to this economic growth. Tea was introduced in Rwanda in 1961, and it is now cultivated on 26,897 hectares by 42,840 farmers spread over 12 of the country's 30 districts, mostly in the country's northern, western, and southern Provinces (Agricultural & Development, 2018). The bulk of tea seedlings are grown on huge plantation areas, with just a modest contribution from tea cooperatives and independent producers. Tea plantations cover the entire undulating hills, their deep green a stark contrast to the blue skies, dirt roads, and sunshine.

Tealeaves are processed in a dozen tea factories located around the country. These factories are available to the public, allowing visitors to learn how tea is gathered and processed, as well as taste the final product. Rwandan tea is grown on hillsides between 1,900m and 2,500m in elevation, as well as on well-drained marshes between 1,550m and 1,800m. Tea production has gradually grown, from 60 Metric Tons in 1958 to almost 30,000 Metric Tons in 2020. Rwanda tea is regarded to be superior due to its great quality, and it is ranked among the best on the planet. Rwanda produces some of the greatest tea grades, including black tea, orthodox tea, white tea, green tea, organic tea, and spicy tea. Rwanda tea is now highly appreciated in the weekly East African Tea Trade Association auctions in Mombasa, fetching record prices in recent years. Its primary markets are in the Middle East, Pakistan, Kazakhstan, and the United Kingdom.

As a result, there is a need to investigate factors associated with tea production in Rwanda in order to gain a thorough knowledge of the decline in output and to forecast tea production for improved planning and stakeholder interventions in tea cultivation. Several different traditional methods and models about tea prediction have been applied in research but they have limited prediction capability. Mutie Silas & Nderu (2017) have conducted a study about Prediction of Tea Production in Kenya Using Clustering and Association Rule Mining Techniques. The main focus of this study was to outset any relationship in tea production of different months of the year, from 2003 to 2015. As result, they found out that in order to enhance tea production, plan for the future production and increase profitability of the ventures, the tea farmers need more to understand the trends in the production, how it is consumed and the process of exportation.

The purpose of this research is to predict tea output in Rwanda using data mining techniques in order to reveal hidden information that different players in this sector in general and tea farmers in particular may utilize in their daily decision-making. For a deeper understanding of data mining, see Jambekar et al. (2018) and Kadlimatti & Saboji (2019), among others.

Several authors have demonstrated how data mining techniques are acceptable tools for

predicting agricultural output and can yield significantly better outcomes than traditional methods. According to the findings in the paper on data mining discussion in the agricultural discipline, Rudyy (2001) describes how data mining may substantially aid in relating the information acquired from the mined data to agricultural yield estimation. This is also supported by Vamanan and Ramar (2011), who claim that a classification approach in data mining can be applied to soil and crop datasets to establish any meaningful association between variables in the dataset. They then used different mining techniques on the identified variables to determine the existence of any meaningful relationships. According to a study conducted by Kodeeshwari and Ilakkiya (2017) on different data mining techniques used in agriculture, data mining plays a significant role in agricultural decision making and problems in the agricultural field can be efficiently solved by using data mining techniques with raw data analysis. Everingham et al. (2016) used random forest regression to predict sugarcane yield in a study where they used simulated biomass from the APSIM (Agricultural Production Systems Simulator) sugarcane crop model, seasonal climate prediction indices and observed rainfall, maximum and minimum temperature, and radiation as inputs to a random forest classifier and a random forest regression.

As pointed out by Linder et al. (2003), it has been demonstrated that Artificial Neural Network may acquire certain capabilities over classical statistical approaches such as diagnosing, classifying, and forecasting. In fact, traditional statistical approaches, such as regression methods and linear discriminant analysis, presume that all inputs are independent and that there is a linear relationship between input and output variables. Artificial Neural Network, on the other hand, do not always assume independence among all inputs and are based on non-linear mathematics.

In their study about crop production prediction using artificial Neural Network, Subash et al. (2020) suggest using a basic neural network model to forecast several elements such as rainfall level, soil composition, weather, and seasonal change values with high accuracy in order to maximize crop production output in India.
It has also revealed that Random forest outperforms standard statistical methods such as logistic regression in some prediction. For example, in their study on pressure ulcer prediction, Song et al. (2021) compared four models: support vector machine, decision tree, random forest, and artificial neural network and found that the random forest model had the highest accuracy for pressure ulcer prediction.

Material and methods

The dataset used in this study was obtained from National Agriculture Export Development Board (NAEB), the government organization in Rwanda in charge of managing exportation of agricultural products. The original identified dataset consisted of 114 observations collected between 2010 and 2019. The features in data set that were used to predict tea production were: year, month rainfall (mm), seedling (seed), fertilizer (Kg) and area under plantation (ha) .The response variable was the production. Data was cleaned removing noise and outliers. Outliers were removed by using z-score method and missing values were handled by imputation. Moreover, cross validation was carried out in order to ease analysis. The 10 K-folds were identified to be more optimal and give better accuracy than others. Cleaning the dataset reduced misclassification and ensured improved model performance. This study used supervised data mining techniques to predict

tea production in Rwanda. Some of data mining models that were used include linear regression (multiple linear regression), K-Nearest Neighbor, random forest and Extra trees.

K-Nearest Neighbor regression

K-Nearest Neighbor (KNN) regression is an instance grounded lazy learning algorithm. It is non-parametric regression model, which does not make any supposition on the distribution of data, thus stimulating training phase. KNN learns complex label function rapidly without losing information. For a given input features x of training set, K observations with x_j in the proximity are considered and the average of the rejoinder of those K predictors (independent variables) gives the predicted output –(Goyal et al., 2014).

Random Forest Regressor

A random forest is a tree-based ensemble, which contains many weak decision tree learners. These weak learners are grown in parallel to minimize the bias and reduce the variance of the model as well (Breiman, 2001). To train a random forest, n bootstrapped sample of datasets are drawn from the novel dataset. Each sample, which has been bootstrapped, is then utilized to grow an un-pruned regression. Instead of utilizing all predictors, which are available in this step, a small and fixed number of K predictors, which have randomly been sampled, are chosen as split candidates. There should be recurrence of these two steps till C trees are grown, and new data is projected by combining the projection of the C trees —(Ahmad et al., 2018). Random forest uses bagging to upsurge the trees diversity by growing them from diverse training datasets, and thus the overall variance of the model is reduced(Rodriguez-Galiano et al., 2015).

Extremely Randomized Trees (Extra trees) Regression

The extra trees regression is as an extension of random forest algorithm and has low chance of over fitting a dataset (Geurts et al., 2006). Extra trees regression utilizes a random subset of input features for training each base estimator just like random forest. Nevertheless, it randomly chooses the feature which is the best along with the conforming value for splitting the node (John et al., 2016). Extra tree utilizes the entire dataset to train each regression tree —(Ahmad et al., 2018).

Evaluation Criterion

In this work, various evaluation metrics were used to evaluate the machine learning models. The used criteria are Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Coefficient of Determination R^2.

Results and discussion

After carrying out data prepossessing and feature engineering, the relevant predictors were identified. Further, cross validation was carried out by splitting the dataset into training set and test set in order to train data mining models by utilizing training set and validate their performance utilizing test set. For training set 80% of dataset were used to ne-tune the algorithms. For the test set 10% of dataset were hold back from training of the model in order to be utilized to evaluate performance of model on unseen data. After training the selected data mining models, their performance was compared by utilizing test data. The initial data mining comparison was done before carrying out hyper parameter tuning (by utilizing default parameters of data mining models) on test set. Later, it was compared after carrying out hyper parameter tuning (by utilizing default parameters of data mining models) on test set. The error metric such as RMSE, MAE and R^2 was utilized to

depict model performance and find the most robust data-mining model.

Table 1 shows comparison of various models on test data before hyper parameter tuning. The error metric such as RMSE, MAE and R^2 was used to find most robust model. It is noted that R^2 was most preferred metric for this study. The R^2 for each of the models were: random forest (0.7894), extra tree (0.7709), lastly linear regression (0.7029) and KNN (0.6734). The RMSE for each of the models were: random forest (1.16E+06), extra tree (1.21E+06), linear regression (1.37E+06) and lastly KNN (1.44E+06). The most robust model before hyper parameter tuning was random forest regression.

From Table 2, the R^2 for each one of the models were: extra tree (0.8953), random forest (0.8340), KNN (0.7302), and lastly linear regression (0.7029). The RMSE for each one of the models were: extra tree (8.15E+05), random forest (1.03E+06), KNN (1.31E+06), and lastly linear regression (1.37E+06). The most robust model before hyper parameter tuning was extra tree regression.

Feature Importance using Extra Tree Regression

Extra tree regression was used to identify important features that contribute to tea production. Season in month(s) was most important feature, followed by rainfall, area under plantation, seedling, fertilizer and lastly year.

This research explained how data mining techniques were used to forecast tea production in Rwanda. It demonstrates that data mining approaches outperform traditional methods in forecasting tea production. This conclusion is not surprising given that several data mining approaches such as linear regression, clustering,

Table 1: Comparison of models before hyper parameter

	Model	MAE	MSE	RMSE	R Squared
1	Linear	1.18E+06	1.89E+12	1.37E+06	0.702852
2	KNN	1.13E+06	2.07E+12	1.44E+06	0.673428
3	Random forest	1.02E+06	1.34E+12	1.16E+06	0.789402
4	Extra tree	1.04E+06	1.45E+12	1.21E+06	0.770851

Table 2: Comparison of models after hyper parameter

	Model	MAE	MSE	RMSE	R Squared
1	Linear	1.18E+06	1.89E+12	1.37E+06	0.7029
2	KNN	1.09E+06	1.71E+12	1.31E+06	0.7302
3	Random forest	9.15E+05	1.05E+12	1.03E+06	0.8340
4	Extra tree	6.06E+05	6.64E+11	8.15E+05	0.8953

association rule data mining techniques, and Random Forest have been shown to outperform in a variety of agricultural domains. These findings of this study were related to previous research findings. For example, in the same way, using clustering and association rule data mining techniques, the findings from the mined data show that average tea output has been the highest trend in the majority of the months in Kenya during the previous couple of years (Kodeeshwari & Ilakkiya, 2017).

According to the findings by Karthigadevi (2020) on the study about Random Forest Classification Algorithm for Agricultural Data Analysis in Tirunelveli District, the Random Forest Classification method provides relevant attributes when compared to the chi-square map reduce, Information gain, chi-square and gain ratio features selection methods and when compared to existing feature selection approaches, the suggested experimental findings demonstrate that the random forest classification algorithm for agricultural data analysis produces high accuracy while taking less time to analyze. This result coincides with our research findings, even though it is about categoriz-ation methods.

The results of a study on crop production prediction using an artificial neural network, in order to maximize agricultural production output in India, Subash et al. (2020) propose utilizing a simple neural network model to anticipate numerous parameters such as rainfall level, soil composition, weather, and seasonal change values with high accuracy. It has also been discovered that Random Forest beats traditional statistical approaches such as logistic regression in some prediction. This finding was revealed in previous studies (Walsh et al., 2017). Moreover, the findings of

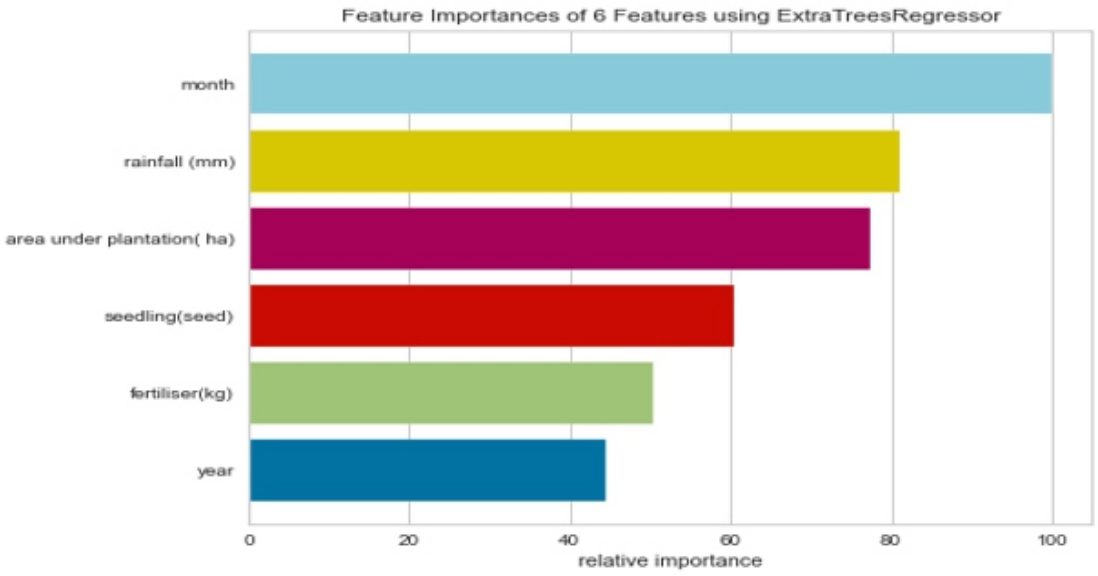

Figure1: Feature importance

a study conducted by ——'Everingham et al. (2016) on accurate prediction of sugarcane yield using a random forest algorithm, indicated that RFR is the best technique.

Tables 1 and 2 show the performance of prediction data mining models as R^2 and error metrics are varied. Using our dataset, it was discovered that altering the error threshold did not result in a substantial improvement in tea production. According to the results in Table 1, random forest outperforms all other models. Because the findings in Table 1 were acquired before the models were modified, they demonstrate that the default parameters for random forest outperform the default values for other models in the data set used in this study. According to Table 2, tree-based models performed marginally better than non-tree-based models. Depending on the situation, the minimal number of samples necessary to divide an internal node might be a significant hyper-parameter.

Despite the fact that similar models were employed, the findings in Tables 1 and 2 differ. After parameter tuning, the model's performance in the table has improved. The use of optimum parameters in Table 2 results in an improvement in R^2 and a reduction in error measures such as RMSE and MAE. The R^2 findings in Table 2 show that all data mining models utilized in this study had a score more than 0.70, which is considered good enough to make a choice. The R^2 for extra tree regression, on the other hand, was spectacular with a score of 0.90, while the least performing model was linear regression (0.70), since the bigger the R^2, the better the regression model in the observations. The explanatory variables can explain 90% of the variance in the true class (tea production) in extra tree regression. In this study, R^2 was preferred because it provides precise predictions when R^2 is high. Extra tree regression might be used to forecast tea production in Rwanda based on R^2 values.

Because of the RMSE and MAE results for data mining models on test data, the results in Table 2 were preferred over the results in Table 1. Table 2 shows that extra tree regression had the lowest RMSE and MAE values when compared to the other data mining models used in this study. According to the results in Table 2, the least performing model in this study was linear regression, which had the highest values of RMSE and MAE. The RMSE and MAE were used to calculate the dispersion of error in the dataset by comparing the predicted and observed values. Based on the RMSE and MAE values, we conclude that extra tree regression is the most robust model and, as a result, can be recommended for predicting tea production using this dataset.

Based on the results in Table 2, extra tree regression, which has the highest R^2 and the lowest RMSE and MAE when compared to other data mining models, was chosen as the best model and was used to find the most important features, as shown in Figure 1. According to Figure 1, the season associated with the months is the most important feature that is highly correlated with tea production. It was discovered that tea production is very high in some months, such as the fifth and eleventh months, while production is low in the eighth month. Rainfall is the second most important factor in predicting tea production. Rainfall, which is part of the weather, is also thought to contribute more to tea production. Tea production increases with an increase in rainfall up to a certain limit (2500 mm in our case), but not beyond that limit. The third most important feature was the area under plantation. It was found that increasing the cropped area also results in increasing tea production. More land for tea plantations might be cultivated in order to increase production. The fourth most important feature

was seedlings, and it was discovered that growing more seedlings ends up with increased tea production.

Conclusion

The feasibility of using data mining models (extra trees, random forest KNN, and linear regression) to predict tea production in Rwanda was assessed in this study. The ability of extra tree and random forest regression to predict tea production has been varied, with the models' prediction accuracy improving. Different metrics of MAE, RMSE, and R were used to evaluate the prediction performance of the data mining models. Extra tree and random forest were found to perform marginally better than the widely used data mining models KNN and linear regression. The study also proposed using extra tree regression to provide insight into the importance of each input feature analysis. The analysis presented here will help researchers and practitioners in the industry gain a better understanding of tea production. The developed data mining models can be used to forecast tea production based on various months, rainfall (climatic conditions), plantation area, number of seedlings sown, and fertilizer type and application rates. The extra tree and random forest regression have the advantage of having only a few tuning parameters and, in most cases, the default hyper-parameter can result in satisfactory prediction performance. Random forest uses out-of-bag samples for internal cross validation and can be applied to any type of dataset. The proposed extra trees algorithm is more computationally efficient and therefore better suited for online or control applications. The developed extra tree model can be used to forecast future tea production. Other data mining models, such as extreme gradient boosted regression, cat boost regression, and logistic regression's performance in the prediction of tea production must be evaluated as well, must be researched as well. Future research will also look at how data mining models perform at different timescales and under different climate conditions. Separate models based on weather classification (i.e., classifying weather based on temperature, clear sky, cloudy day, foggy day, etc.) will also be investigated in the future. Exploration of Big Data technologies for training and deploying prediction models is also required.

Separate models based on weather classification (i.e., classifying weather based on different weather conditions such as temperature, clear sky, cloudy day, foggy day) should be investigated in the future, according to this study. We also recommend that the government investigate other underlying factors that may have influenced tea production during different months (seasons) but were not captured in the provided dataset. For example, it was discovered that tea production was high in some months and low in others. Aside from the suggestions made in this study, there is a need to identify additional factors that can be identified through the use of sensors and other data mining tools. The authors recommend using data mining models to predict tea production, particularly the extra tree and random forest models, which were found to be more effective in this study. The authors also advise the Rwandan government to look into Big Data technologies in order to fine-tune and deploy prediction models, particularly for forecasting tea production.

Acknowledgements

The authors acknowledge NAEB for granting access to the data, and African Centre of Excellence in Data Science (ACE-DS) for funding this research.

References

Agricultural, N., & Development, E. 2018. NAEB 2017-2018 ANNUAL REPORT. August, 153.

Ahmad, M. W., Reynolds, J., & Rezgui, Y.

2018. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. Journal of Cleaner Production. https://doi.org/10.1016/j.jclepro.2018.08.207

Breiman, L. 2001. Random forests. Machine Learning. https://doi.org/10.1023/A:1010933404324

Everingham, Y., Sexton, J., Skocaj, D., & Inman-Bamber, G. 2016. Accurate prediction of sugarcane yield using a random forest algorithm. Agronomy for Sustainable Development, 36(2). https://doi.org/10.1007/s13593-016-0364-z

Geurts, P., Ernst, D., & Wehenkel, L. 2006. Extremely randomized trees. Machine Learning. https://doi.org/10.1007/s10994-006-6226-1

Goyal, R., Chandra, P., & Singh, Y. 2014. Suitability of KNN Regression in the Development of Interaction based Software Fault Prediction Models. IERI Procedia. https://doi.org/10.1016/j.ieri.2014.03.004

John, V., Liu, Z., Guo, C., Mita, S., & Kidono, K. 2016. Real-time lane estimation Using Deep features and extra trees regression. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-319-29451-3_57

Karthigadevi, K. 2020. Random Forest Classification Algorithm for Agricultural Data Analysis in Tirunelveli District. Journal of Xian University of Architecture & Technology, XII(VIII), 418432.

Kodeeshwari, R. S., & Ilakkiya, K. T. 2017. Different Types of Data Mining Techniques Used in Agriculture - A Survey. International Journal of Advanced Engineering Research and Science, 4(6), 1723. https://doi.org/10.22161/ijaers.4.6.3

Linder, R., Mohamed, E. I., De Lorenzo, A., & Pöppl, S. J. 2003. The capabilities of artificial neural networks in body composition research. Acta Diabetologica, 40(SUPPL. 1). https://doi.org/10.1007/s00592-003-0018-x

Mutie Silas, N., & Nderu, L. 2017. Prediction of Tea Production in Kenya Using Clustering and Association Rule Mining Techniques. American Journal of Computer Science and Information Technology, 05(02). https://doi.org/10.21767/2349-3917.100006

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews. https://doi.org/10.1016/j.oregeorev.2015.01.001

Rudyy, R. 2001. On the classification of agricultural lands. Archiwum Fotogrametrii, Kartografii i Teledetekcji, 11(1), 379384.

Subash, M., Sawant, S., Kishore, V., & Sandhya, P. 2020. Crop Production Prediction Using Artificial Neural Network. *Journal of Critical Reviews*, *7*(17), 2064-2072.

Linder, R., Mohamed, E. I., De Lorenzo, A., & Pöppl, S. J. 2003. The capabilities of artificial neural networks in body composition research. *Acta diabetologica*, *40*(1), s9-s14.

Song, J., Gao, Y., Yin, P., Li, Y., Li, Y., Zhang, J., & Pi, H. 2021. The Random Forest Model Has the Best Accuracy Among the Four Pressure Ulcer Prediction Models Using Machine Learning Algorithms. *Risk Management and Healthcare Policy*, *14*, 1175.

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. 2017. Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science*, *5*(3), 457-469.