International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Application of Random Forest in Campus for Attendance Predictions

Jahanvi Mehariya[1], Chaitra Gupta[2], Niranjan Pai[3], Sagar Koul[4], Prashant Gadakh[5]

[1-4]*Student, International Institute of Information Technology*
[5]*Professor, Dept. of Computer Engineering, International Institute of Information Technology, Maharashtra, India*

Abstract - *In large campuses with various courses, it is often seen that attendance of students is a major concern. Universities spend a lot on the infrastructure like classrooms, labs and halls so as to accommodate all the students based on the enrollment numbers but more than half of the space remains unused, this is because the actual attendance is way less than enrollment. This is due to many reasons like these days online content is available so students don't feel the need to attend. This system aims to achieve an optimal room occupancy by allocating classes based on the actual attendance of the students. This will solve the problem since time-table and allocation can be well planned if count of actual students who will attend the class will be known in advance. The model is trained to predict attendance by using random forest algorithm. It learns the pattern of student attendance from historic data & then based on the learning it will predict how many students will attend a lecture at time abc of subject xyz on last day of the week and so on. Regressor will be used as output is of continuous type i.e. numerical value and not binary outcomes. Predictor variables are year, semester, week, day, date, time_of_day, start_time, end_time, room_name, class_type, faculty, joint, school, status, degree, enrollment, class_duration and output is attendance.*

Key Words: **Predictor Variables, Random Forest, Regressor, Root Mean Square Error**

1. INTRODUCTION

Campuses have various courses and attendance of students is a major concern especially in large campuses. They spend a lot on the infrastructure but most classroom spaces remain under-utilized. Students enroll in huge numbers but when semester begins students actual attendance is very less. So, this algorithm learns pattern from historic data and tries to predict the attendance of students given a particular day of week, subject etc. This prediction will help to know the attendance of students beforehand which will make planning of time-table and allocating classrooms based on that attendance and campus infrastructure cost can be optimized.

In this paper we have considered University of New South Wales Smart Campus data from smartcampus.unsw.edu.au website [3]. The model is trained using this dataset by splitting the data into training and testing. The algorithm applied is random forest regressor. This is the first module which is data pre-processing, cleaning and training the model to give accurate predictions. Random forests are also called as decision forests as they are built from them. There are different types of learning algorithms supervised, unsupervised and reinforcement. Random forest falls under supervised type of learning.

Random Forest is the most accurate in predicting the attendance. Some predictor variables do not contribute much to the attendance output and this can lead to overfitting, one way to avoid this is choosing random forest as the technique. The fitted random forest regressor model can be used to predict the future attendance. This model gives very accurate results.

2. LITERATURE SURVEY

We have referred "Experiences with IoT (Internet of Things) and AI (Artificial Intelligence) in a smart campus" as base paper [5] in which they have implemented a system for dynamic allocation of classes. In the first module sensor-based techniques for gathering the actual count of students in class are compared. Beam counters, wi-fi, thermal sensors, cameras are used. In the second module algorithms for training model to predict future attendance are compared. Multiple regression, support vector machines and random forest were used. We have chosen random forest as it produced the most accurate result. It has the least RMSE (Root Mean Square Error) value. Many factors were considered as input to the model like subject name, undergraduate or post graduate course, time of the lecture, day of the week, duration of the lecture, faculty, type of lecture i.e. theory or practical, etc. The output is the count of students who will be attending the lecture.

3. RANDOM FOREST REGRESSION

Regression is used to understand the relationship between one or more input variables and an output variable. We want to understand the impact of inputs like time of the lecture, day of the week, type of lecture whether theory or practical on the output that is whether the student will be attending the lecture or not. Since the output of our model is a continuous variable we have used regression and not classification.

Random forest is built by taking random samples from dataset and building decision trees and then merging these decision trees into a forest. The goal is to achieve better accuracy as the model does not rely on a single decision tree but multiple ones.

Random forest is a bagging technique since while building the trees they are built independently as they run in parallel.

- Feature Importance: It is very easy to measure the relative importance of each feature on the prediction. Feature importance is needed to reduce overfitting as by seeing the importance of features we can decide whether that feature contributes or not to the output and if it contributes but it hast very less percentage of impact on the output then it can be discarded.
- Gini Index: Is the measure of impurity. The attribute with least impurity is selected as best attribute.
- Entropy: is the degree of randomness or impurity in the dataset. It should be low.
- Information gain: is the entropy of the parent node- the entropy of all child node. It should be high.

These methods are calculated to decide the node to be used for splitting.

Algorithm of the Random Forest:

- First, randomly select few samples from the training set.
- Next, construct decision tree for every sample. Each decision tree will predict a certain result.
- In the next step, voting will be performed for every predicted result.
- At last, select the result which gets the maximum number of votes and assign this result as the final predicted output

Advantages of Random Forest:

- Random forest produces very accurate results since it takes the results from various multiple decision trees.
- The algorithm can be used for both classification and regression problems. That means whether the output is binary yes-no or 0/1 or the output is continuous i.e., numerical type we can classifier or regressor repectively.
- Random forests can handle missing values. There are two ways to handle these: using median values to replace continuous variables or computing the proximity weighted average of the missing value.
- It also helps for feature selection, hence reducing overfitting.

Limitations of Random Forest:

The main drawback is that the greater the number of decision trees used to build the forest, more processing time is consumed. So where run-time is given more preference over accuracy Random Forest is not used. It is not time-efficient algorithm.

Accuracy:

Comparing the result from a particular row of inputs of the dataset with the result we obtained by the model after training is quite high as the predicted attendance is 135 and the actual attendance is 133 which is quite close.

```
test=np.array([2017,0,4,33,2,0,5,7,6,0,6,12,0,4,2,291,1])

y_pred1=regressor.predict([test])

y_pred1

array([135.4])
```

Fig -1: Output

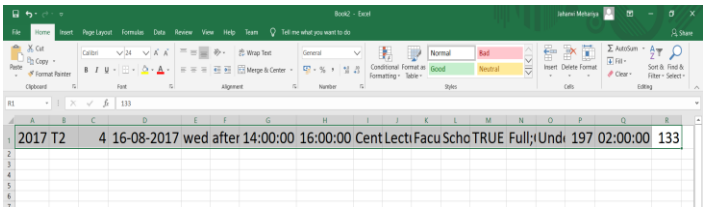

Fig -2: Dataset Entry

In train.csv the last record has attendance as 133 and model predicted it as 135 so accuracy is high.

Steps:

1. Import pandas and other libraries in python and load the dataset into a dataframe
2. Clean the dataset by using LabelEncoder and normalize the attributes.
3. Split the dataset into training and testing set by choosing an appropriate split ratio
4. Create random forest regressor model using Sklearn packages and fit the training data by specifying number of decision trees i.e. n_estimators
5. Visualizing the Random Forest Regression results

Output values are numerical. Attendance is the predicted result. It is also assumed that the training data is independently selected from the original dataset.

Evaluation metrics for regressor models:

Root Mean Square Error: The accuracy of regressor model is measured by its RMSE value. This is the average of the difference between the actual output and the output predicted by the model RMSE is square root (mean squared error)/100 RMSE=sqrt(mean((observed-predicted)^2)/number of cases). The lower the Root Mean Square Error, the more accurate the model.

Mean Squared Error (MSE): MSE is the average of squares of the “errors”.
Mean Absolute Error (MAE): MAE is the difference between two variables which are of continuous type.
Error is the difference in predicted value and the actual output.

Visualization: A heat map represents the correlation between predictor variables and any similarity between represented using different colors.

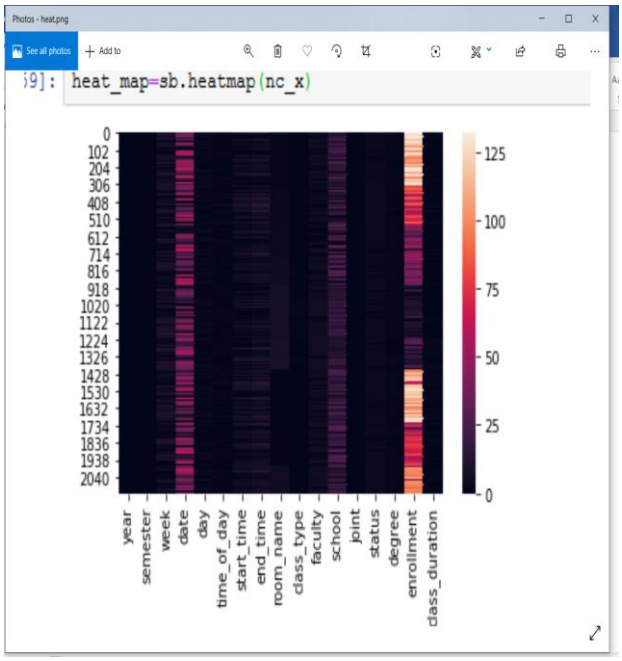

Chart -1: On x-axis the input features or predictor variables have been plotted and on y-axis the similarity between them has been represented.

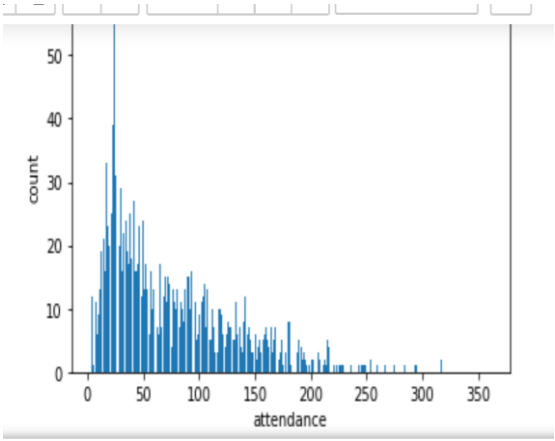

Chart -2: Bar plot shows the variation of attendance in the dataset.

Output:

```
from sklearn import metrics

print('Mean Absolute Error: ',(metrics.mean_absolute_error(y_test,y_pred)))
print('Mean Squared Error: ',(metrics.mean_squared_error(y_test,y_pred)))
print('Root Mean Squared Error: ',(np.sqrt(metrics.mean_squared_error(y_test,y_pred))))
```

```
Mean Absolute Error:  0.08743261682242992
Mean Squared Error:  1.860927415233644
Root Mean Squared Error:  0.13641581342475087
```

Fig -3: Code

We get an RMSE of 0.12 and the models in the base paper have RMSE in range of 0.12 to 0.16.

Dataset	Source	Input Attributes	Output	Total
UNSW Smart Campus data	smartcampus.unsw.edu.au	year, semester, day, date, week, joint, faculty, school, status, degree, enrollment, room_name,class_type, time_of_day, start_time, end_time, class_duration	Attendance (numerical)	Train (2137) Test (1043) Total (3180)

Below is a table representing the dataset:

Table -1: Sample Table format

4. CONCLUSION

In this paper we consider "University of New South Wales Smart Campus data" from smartcampus.unsw.edu.au website and applied random forest algorithm with help of Sklearn using python programming language and achieved an RMSE of 0.12, giving better results than other techniques like multiple linear regression, support vector machines, which produced an RMSE above 0.16. Hence, proving that random forest gives the best results. Further we can apply this random forest algorithm for other data sets as well and also we can take variety of other inputs affecting attendance of students like location of students whether they are hostelites or localites, even events and weather conditions like rainy season can be taken into consideration.

REFERENCES

[1] S. Devadoss et al., "Evaluation of factors influencing student class attendance and performance," American Journal of Agricultural Economics, vol. 78, no. 3, pp. 499–507, 1996.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2] L. Breiman et al., "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.(2018) UNSW Smart Campus - Classroom Usage Monitoring and Optimization. https://smartcampus.unsw.edu.au/roomoccupancy/data/IoT-J/.

[3] Jitendra Kumar Jaiswal, Rita Samikannu, "Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression," IEEE, 978-1-5090-5573-9/16, 2017

[4] Thanchanok Sutjarittham, Hassan Habibi Gharakheili, Salil S. Kanhere, and Vijay Sivaraman, "Experiences with IoT and AI in a Smart Campus for Optimizing Classroom Usage," IEEE, 2327-4662 (c), 2018

[5] T.Sujaritham et al., "Data-Driven Monitoring and Optimization Classroom Usage in a smart Campus ," in Proc.ACM/IEEE IPSN,Porto,Portugal,2018.

[6] J.Li,L Huang, and C.Liu,"Robust people counting in video surveil-lance:Dataset and system,".inProc.IEEE AVSS,klangerfurt,Australia,2011.

[7] K. S. Liu, J. Francis, C. Shelton, and S. Lin,"Long Term Occupancy Estimation in a Commercial Space:An Empirical Study"in IEEE IPSN,Pittsburg, USA, April 2017.

[8] "School Management System ", Iolite.org.in, 2018. Available:http://www.iolite.org.in./.

[9] A.Dammak, A.Elloumni and H.Kamoun." Classroom assignment for exam timetabling,"Advances in Engineering Software,pp659-666,2006.

[10] P.KumarChaki and S. Anirban," Algorithm For Efficient Seating Plan For Centralized Exam System", International Conference on Computational Techniques in Information and Communication Technologies, 2016.

[11] D. Amaratunga, J. Cabrera, and Y.S. Le,." Enriched random forests. Bioinformatics", 24:2010–2014, 2008.

[12] (2017) EvolvePlus: Overhead Thermal Counter. https://goo.gl/ugtW1L.

[13] (2017) Steinel Australia : Wireless People Counter with Remote Data View-ing. https://goo.gl/AUe8G

[14] Andy Liaw and Matthew Wiene, "Classification and Regression by randomForest", Vol. 2/3, December 2002