# Project Overview

Environment is a very essential thing for a person to work. If a person is working in a suitable place, then he or she can focus on their work properly. In the bigger workplaces employees are assigned with individual rooms which they must use. In this case, there can be situation where a person can get a great room which has a window and gets cool wind to work in a peaceful environment. On the other hand, there can be another employee who gets a room where there is too much warmness, and it harms his work. Sometimes it can harm that person’s mental health as well. To choose a room there can be many factors such as if that room has proper humidity if light properly enters that room and many more. Based on those factors if a choice is given to a man that if he or she will use that room or not then they can come to a decision. Our project is about this room occupancy factor. In an office there are many rooms and those room has some factors such as humidity, wind, light etc. and based on those if that room will be occupied or not. This model will be useful in the future for a office to provide proper rooms to their employees. This project is a bit different than those traditional projects such as cancer detection and many other medical classification models. This project is based on a real-life dataset and as a result, its outcome can be applied in the real-life problems. Moreover, this dataset has thousand of instances so it can be said that its outcome and accuracy can be classified as concrete. This project has the target audience of the startup companies because as a new company they often are unable to provide their workers with a proper workplace. As we have already told this project is different than other normal models, because in real life we saw our own people struggling in their office without proper workplace. From that, we got the inspiration that we want to make a classification model based on room occupancy. So that, in the future this problem can be reduced as much as possible. Over the next pages we will discuss about the project model, its accuracy and compare between different model to find the best results. Now comes the question how we will test the dataset and make the model. We will use the graphical user interface tool named WEKA to evaluate everything. It has everything to complete this project with proper visualization tools as well.

# Dataset Overview

To make a classification-based model we need a proper dataset and as of requirement we needed a dataset with at least 1000 instances, and it should be real.

Dataset Name: **Room** **Occupancy Detection Data Set**

Link: **http://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+**

This dataset is a classification-based dataset which is based on real data. Which means our model will have a real-life impact. It has total 8143 instances or row of data. It has total 7 columns or attributes. From them 5 are important and can be called the feature matrix and the last one is the class attribute which is used for prediction. Our dataset is about predicting in a office will a certain room be occupied or not based on some parameters or attributes. Now let us know about those parameters and why they are important for a room to be occupied or not.

**Temperature:** Our first attribute is temperature which is given in Celsius. When you choose a room always you want a room not to be too hot or cool. If a room always has high temperature or very low temperature it will not be suitable for work.

**Humidity:** Humidity is also another factor to choose your room in the office. When the humidity is between 30 to 50% it is better. If the humidity is high the outside feels wetter and if it is higher than 60% it will cause health issues.

**Light:** Light can be said as the most important factor for a room to be occupied or not. If a room does not have proper lighting, then there is a very high chance an employee will not use that room for work.

**CO2:**  CO2 Level is also an important factor for a room. High CO2 levels indicate a potential problem with air circulation and fresh air in a room or building. It should be lower than 1000 ppm in all the case for a room to be in fresh condition.

**Humidity Ratio**: It is also known as the Specific or Absolute Humidity, the finite quantity of moisture in each volume of air. It should be well balanced in a room otherwise a room will not be suitable for working.

# Model Development

**NAÏVE BAYES**

The K-Fold Cross Validation has been used where K value is 10.

|  |
| --- |
| Figure : Each Attribute’s Mean, Standard Deviation etc. |

In this figure we can see each attribute’s standard deviation value, mean, weighted sum and average. As all the feature columns are numeric so these values can be figured out and they play a role in the prediction.

|  |
| --- |
| Text  Description automatically generated  Figure : Confusion Matrix and Predictive Accuracy of the model |

In figure 2, the confusion matrix of the naïve bayes model can be seen also with the model’s

predictive accuracy which is very high 97.7%. Considering the factor that naïve bayes only works with categorical data this model has given very high accuracy with being only 187 instances being wrongly predicted.

**K-Nearest Neighbor**

Here is also the K-Fold Cross Validation Technique is used for evaluating the model and K value is considered as 3.

|  |
| --- |
| Graphical user interface, text, application, email  Description automatically generated  Figure : Initial Information’s about the dataset and evaluating method. |

In the figure 3, the initial information’s about the evaluation of this dataset is described. It contains all the feature and class attribute the training method and the nearest neighbor value which is 3 in this case.

|  |
| --- |
| Text  Description automatically generated  Figure : Confusion Matrix and Predictive Accuracy for KNN |

In the figure 4 the confusion matrix of k nearest neighbor classifier and predictive accuracy of this model can be seen. As KNN better works with the numeric values and this dataset is

mainly based on numeric values this classifier is great suite for this dataset and this accuracy reflects it. With 3 nearest neighbor values the accuracy is 99.3% which is very high considering this dataset having more than 8000 instances.

**Decision Tree**

Here the J48 variant of decision tree is used, and the evaluation has been done using the 10-Fold Cross Validation. Below you can see the pruned version of the whole decision tree using j48.

|  |
| --- |
| Figure : The Full Decision Tree for this Model |

In this figure 5 we can see the whole decision tree of j48 variant for our dataset. It has total 22 leaves and size of the tree is 43. The root of this tree is the light attribute. As we discussed in the dataset overview section it was considered as the most valuable attribute and as a result it is also in the root of the tree. The leave is as expected contains the class attribute.

|  |
| --- |
| Text  Description automatically generated  Figure : Confusion Matrix and Predictive Accuracy for J48 |

In the figure 6 the confusion matrix and predictive accuracy for decision tree classifier can be visualized. It has a very high accuracy of 99.5%. As decision tree can work with both categorical and numeric data this accuracy was expected from this model. This model also proves this dataset can be used for real-life application.

# Discussion and Conclusion

**COMPARISON OF MODEL USING CONFUSION MATRIX**

|  |  |  |
| --- | --- | --- |
| Figure : Naïve Bayes | Text  Description automatically generated with low confidence  Figure : KNN | A picture containing table  Description automatically generated  Figure : Decision Tree |

In this project we have used three data mining algorithms to evaluate our real-life dataset and their confusion matrix is listed upwards. As we got to know the naïve bayes gave us the lowest accuracy among them which was 97.7%. On the other hand, both KNN and Decision Tree gave us very high accuracy of over 99%. The False Negative/Positive value ratio is very low in these two algorithms compared to the naïve bayes algorithm.

**Conclusion**

This project was about if a certain office’s room will be occupied or not based on some important parameters, and it was based on real dataset. Here the algorithms Naïve Bayes, KNN and Decision gave the predictive accuracy of 97.7, 99.3 and 99.5% accordingly. We can decide that for this dataset decision tree is the most suitable data mining algorithm to apply. But here are some factors, for the KNN, the nearest neighbor value was used as 3. If lower values are used for that then it that case KNN comes out on top. Now let us look at naïve bayes algorithm. We know that naïve bayes best works with the categorical based dataset whereas our whole dataset was numeric, and its only class variable was categorical. Still this algorithm gave us over 97% accuracy. For this reason, we can also say naïve bayes is not a bad algorithm to apply here. Overall, to draw the conclusion we will choose decision tree as the most suitable algorithm.