**IMPLEMENTATION**

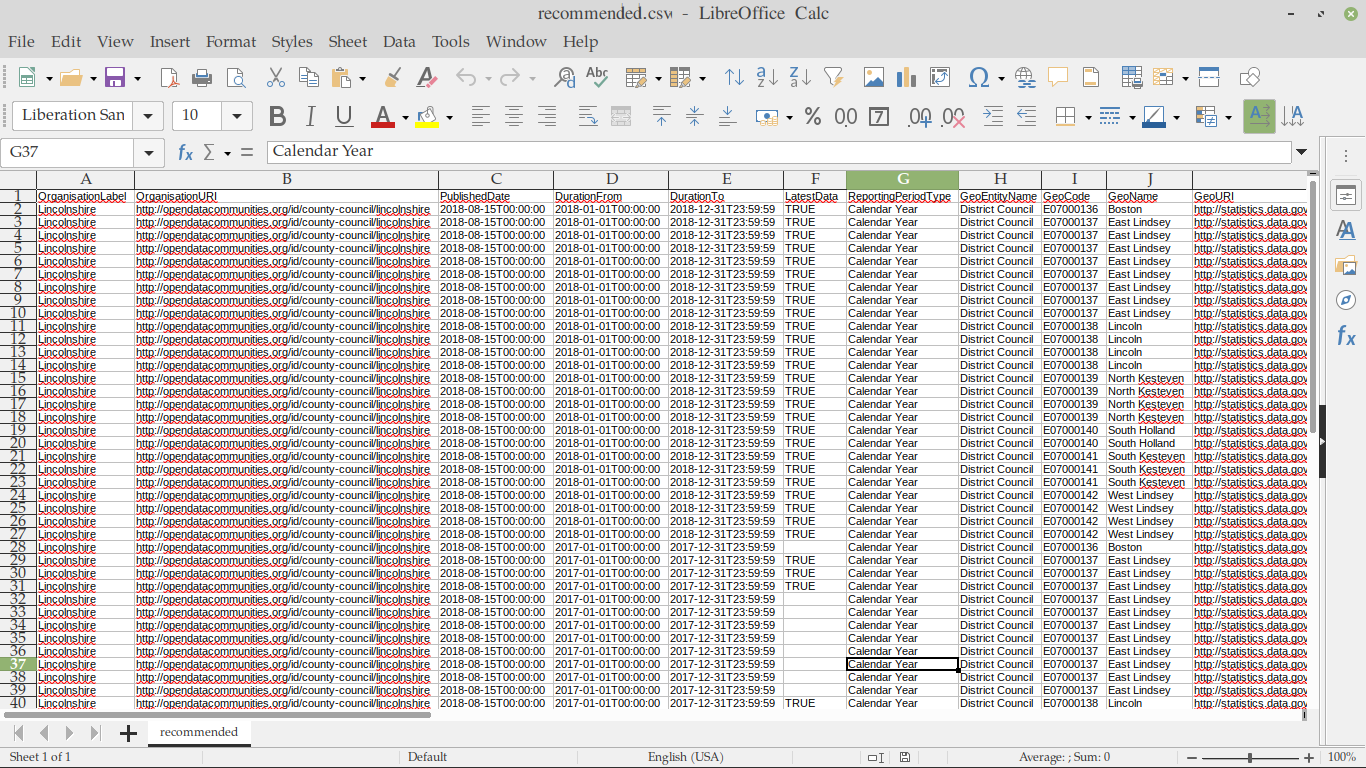
**Introduction**

Every day data increases and becomes more overwhelming by the day. There is a growing gap between data and the understanding of it. As data increases the number of users who understand the data decreases. Lying under this data is potential information that is a useful resource. Therefore, one can check for patterns in data and evaluate the probability of future events. [4]Long before the term data mining had been introduced in the 1900’s, finding patterns from data had long existed. Early data pattern identification techniques include the Bayes theorem (1700s) and regression evolution (1800s). As data sets have a broad size and level of complexity, the generation and increasing power of computer science has boosted data collection, storage, and manipulation. Data collection has been boosted by the growing power of computers as data sets have become broad and in size and the complexity level has broadened as well. Hands on data investigation has progressively improved with indirect, automated data processing and other computer science developments such as neural networks, clustering, generic algorithms(1950’s), decision trees(1960s), and supporting vector machines (1990s).[3] Data Mining is the process of finding potentially useful patterns from huge data sets.[8]The insights from data mining are used in making decisions, analysis on market basket, production control, customer retention, scientific discovery and education systems. [9]Predictive analysis as a feature of data mining analyzes current and historical facts to make predictions about future or otherwise unknown events. [1]Over time, time series analysis is one of the analytic techniques that analyses time series data in order to extract meaningful statistics and other characteristics of the data. Auto-correlation patterns and seasonality measures found in time series models can be used to predict when a certain data point is expected. This has contributed to its use is many fields over the years such as econometric, mathematical finance, weather forecasting, earthquake prediction, control engineering and astronomy, helping this fields make better future decisions. Like many countries, Tourism in Kenya is a big economy and sometimes it fails to produce as much revenue because the country has not studied patterns in the sector well enough to determine the where, why and how to maximize on this national treasure. [6]One of the goals of the Kenyan Tourism Board is to provide quality service delivery. Data mining is a solution in this case to forecast on huge amounts of data to make better business decisions and actionable strategies that will impact profitability in this sector..

**Dataset**

The dataset contains attributes which are in form of a csv file. The dataset is from Data world repository(Kaggle) and is called Tourism – Visitors to Attraction dataset. In order to have consistent data we will pre-process the dataset so as to eliminate the missing and redundant records. These will be used for validation and testing purposes. Below is a sample of the dataset: After data is collected it’s pre-processed to attain its optimal quality which is non-erroneous. Data is then cleaned for better classification purposes. The attributes should not contain any missing values or inconsistencies such as noisy data because if the classification is done with the errors the output will not be true.

This dataset shows how many people visited Attractions in Lincolnshire by calendar year. Visitor numbers for a wide range of attractions are shown, along with other key information such as entrance fees. The data's source is the Annual Survey of Visits to Visitor Attractions, run by Visit England. (As usual with survey data there are some limitations, such as not all visitor attractions participating in the survey, and where visitor numbers are estimated that is indicated in the data). This dataset is updated annually from statistics published by Visit Britain, see the Source link for more information.



**Machine Learning Algorithms**

Classification in machine learning is important as it predicts modelling problems where class label is predicted for a given example of input data. This data-set has been categorized into 2 distinct classes I.e Recommended sites and Not Recommended sites.

**Decision Tree Algorithm**

This is a classification algorithm under supervised learning that can be used to solve both regression and classification problems.

It’s called a decision tree as it starts with the root node that further expands into branches forming a tree like structure.

In a decision tree:

The algorithm begins at the root node which contains the complete training dataset. Then the best attribute in the dataset is found using Attribute Selection Measure techniques like information gain technique or the Gini index.

The root node is then split into subsets that contain the possible values for the best attributes from this decision nodes (which are nodes that can be used to make any decision and can have multiple branches) which contain the best attributes are generated. The decision nodes are recursively split using the subsets until a stage is reached where you cannot further classify the nodes and reach the final leaf node (a leaf node is a node that cannot be further divided).

**Tools Used**

In this project tools used are Anaconda to check attributes and outliers.

Standard libraries for model creation are used. The following are some of the libraries could be used in this system

**Tkinter**

It’s a standard GUI library of python. Python when combined with tkinter provides fast and easy way to create GUI. It provides powerful object-oriented tool for creating GUI.It provides various widgets to create GUI some of the prominent ones being:

* Button
* Canvas
* Label
* Entry
* Check Button
* List box
* Message
* Text

Using tkinter we are able to create an interactive GUI for the model.

**Pandas**

It is the most popular python library used for data analysis. It provides

highly optimized performance with back-end source code purely written in C or python. Data in python can be analysed with 2 ways.

* Series
* Dataframes

Series is one dimensional array defined in pandas used to store any data type. Dataframes are two-dimensional data structure used in python to store data consisting of rows and columns. Pandas dataframe is used extensively in this project to use datasets required for training and testing the algorithms. Dataframes makes it easier to work with attributes and results. Several of its inbuilt functions such as replace were used in the project for data manipulation and processing.

**TensorFlow**

TensorFlow is an open-source framework that comes in handy for large-scale as well as numerical ML. It is a blender of machine learning as well as neural network models. Moreover, it blends with Python. The most prominent feature of TensorFlow is, it runs on CPU and GPU as well. Image classification is highly implemented using tensorflow for optimum accuracy. Its to this reason that this project best suits the use of tensorflow.

**Data Processing**

Data is likely to be noisy, inconsistent and incomplete or have missing information. Such kind of data can highly affect the accuracy of data mining results and hence this calls for pre-processing or rather preparing and transforming the data so as to achieve quite accurate results. This therefore makes data pre-processing a vital step in data mining. Data Processing has various methods and it is divided into four methods which are:

* Data cleaning.
* Data integration.
* Data Transformation.
* Data reduction.

**Data Cleaning**

Data contains missing values meaning it is incomplete this may happen when data was collected in a hurry or some respondents refused to give out some details, noisy in that it has errors or outliers and such data causes confusion and has duplicates. Data cleaning cleans data by doing away with outliers, handling missing values and dealing with inconsistent data. The dataset has been fine-tuned to avoid data inconsistency and does not contain missing data.

**Data Intergration**

Data integration refers to combining data from multiple places and putting it to one central place. This can however be a challenge since we not sure of some things hence data integration was not conducted in the dataset was complete and it contained all the attributes hence no need arises to combine the data.

**Data Transformation**

Data transformation is the most difficult part in preparation of the dataset

in order to produce a data-set that can be used for training of the model.

Here the the data is arranged into appropriate ways to allow effective data mining.

**Data reduction**

Data reduction is helpful since it reduces the dataset volume and still gives a dataset with useful knowledge. Mining on the reduced dataset gives more accurate values. One way to reduce that number (and hopefully produce a more accurate model) is to effectively compress the image. We do this by trying to find a way of keeping as much information as possible about the data without losing the essential structure.

**Training**

In order to successfully implement a ML algorithm, data had to be separated between training and testing. This is because while working with datasets, a machine learning algorithm works in two stages the testing and the training stage. In this project the data was split between test-train 20 percent 80 percent manually where we gave a training dataset containing 80 percent of training data and 20 percent of testing data. For this process to be successful pandas were used to import the dataset and TensorFlow library was used for training function. OS library is used to enable transversion through directories.

RANDOM library enables the shuffling of data during training. Matplotlib is very essential as it help in visualization of the data through

a graph. The classes are put in an array CATEGORY [ ], where Recommended and NOT Recommended are the classes. The label and features are appended to x\_train and y\_train array. The model is created using decision tree algorithm.

Notice that recommended dataset has 40 rows of data, whereas Not Recommended dataset has 46, 20 row have set aside for testing which is 20 percent of the original data to enhance exactly the result we wanted.

**Testing**

The test set is generally what is used to evaluate competing models. It

contains carefully sampled data that spans the various classes that the model would face, when used in the real world.

**Evaluation Metrics**

The test set is generally what is used to evaluate competing models. It

contains carefully sampled data that spans the various classes that the model would face, when used in the real world.

**Accuracy Test**

The accuracy of a machine learning classification algorithm is one way to measure how often the algorithm classifies the point correctly. Accuracy is the number of correctly predicted data points out of all the data points. More formally, it is defined as the number of true positives and true negatives divided by the number of true positives, true negatives, false positives, and false negatives. A true positive or true negative is a data point that the algorithm correctly classified as true or false, respectively. A false positive or false negative, on the other hand, is a data point that the algorithm incorrectly classified. Accuracy of different algorithms in this project were determined using confusion matrix and accuracy score

**Accuracy Score**

Accuracy classification score in multilabel classification, this function computes subset accuracy: the set of labels predicted for a sample must exactly match the corresponding set of labels in y true. Classification accuracy is our starting point. It is the number of correct predictions made divided by the total number of predictions made, multiplied by 100 to turn it into a percentage. The following is an accuracy score from the decision tree prediction: 0.7099 \* 100 =70.99 percent

**Discussion of the results**

We set out to create a model which can recommend the most best places to visit. Such a system can decrease the confusion and improve decision making in travelling. We were successful in creating such

a system where on an average we achieved accuracy of 71 percent accuracy. Such a system can be largely reliable. Creating this system, we also added a way to store the data entered by the users in the database which can be used in future to help in creating better version of such system. Our system also has an easy to use interface.

**Conclusion and Future Work**

Prediction is very difficult especially about the future. In this project there was the need to design a predictive model for traveling areas recommendations with more improvements in accuracy using various data mining techniques on a dataset collected from Kaggle. Performance of the models was evaluated using metrics of accuracy score so as to achieve results that are more accurate and have less variance. The researcher took more time on data processing especially data transformation and dealing with filling in missing data and data preprocessing which is an important step so as to achieve

accurate results. The models performed well on predicting different recommedations.

This paper has brought to the light risk factors of different traveling areas recommendation by data mining techniques. The outcome of this study can be used as an assistant study by other researchers to help them come up with more consistent predicting methods. Most of the experiments done in this study used decision tree algorithm provided under jupyter notebook in the anaconda navigation platform, further investigations should be done with different platforms so as to enhance and expand the capabilities of prediction models. As future work, the researcher has planned to perform additional experiments with more datasets and algorithms to build models that can predict various and more areas but to time limitation, the following research/ work needs to be performed in the future.